

A modeling-based evaluation of the evidential basis for and
cost effectiveness of intensive post-diagnosis extra-colonic
surveillance of non-metastatic colorectal cancer patients

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE
UNIVERSITY OF MINNESOTA
BY

JONAH POPP

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Advisor: Karen Kuntz

March, 2018

© Jonah Popp 2018

Acknowledgements

I would like to thank my advisor Professor Karen Kuntz for her mentoring, feedback, and support throughout my doctoral program. I would also like to thank my other committee members for their support and feedback on my work: Professor John Nyman, Dr. Eva Enns, and Dr Xianghua Luo.

Abstract

Between 70-80% of colorectal cancer (CRC) patients present with non-metastatic disease and can potentially be cured with surgical resection. However, between 5-60% of these patients will suffer a recurrence, generally in the form of late-occurring metastatic disease. For this reason, most professional-society guidelines recommend intensive extra-colonic-focused surveillance (CT scans and routine testing for tumor-markers) of these patients for 3-5 years post-diagnosis with the aim of detecting recurrence at an earlier stage when it is more likely to be amenable to salvage surgery with a curative intent. Until recently, this practice was corroborated by the results of meta-analyses of randomized control trials (RCTs) comparing more intensive with less intensive (or no) surveillance. However, the negative results of two large recently-published RCTs – the UK FACS trial and the Italian GILDA trial - and of subsequently updated meta-analyses have cast doubt on the value of aggressive follow-up and ultimately the value of aggressive treatment of recurrent CRC.

In this dissertation I use a modeling analysis to argue that the results of these two trials have been misinterpreted. Accordingly, the conclusions of the most recent meta-analyses are misguided and calls to throw in the towel on intensive follow-up are premature. The negative trial results are not surprising given the low recurrence rates of contemporary practice and thus the small proportion of patients who could potentially benefit from aggressive follow-up. I show that, if aggressive follow-up were to confer a survival advantage in virtue of increasing the chances of salvage therapy with a curative

intent, the average benefit would be very small. Moreover, the two trials would have had essentially no chance to detect an effect of that size, and this problem of insufficient power was likely exacerbated in at least one of the trials by a sizable chance recurrence imbalance. I further show that it is unlikely that a RCT with adequate power could ever or will ever be possible. However, I argue there is reason to take seriously the hypothesis that aggressive use of follow-up testing and subsequent salvage therapy can offer a small survival advantage on average. Finally, I report the results of a modeling-based cost-effectiveness analysis to identify follow-up strategies that would be cost-effective if this hypothesis is correct.

Table of Contents

List of Tables.....	v
List of Figures.....	vii
Chapter 1.....	1
Chapter 2.....	15
Chapter 3.....	62
Chapter 4.....	97
Bibliography.....	151
Appendix A.....	167
Appendix B.....	174
Appendix C.....	195
Appendix D.....	197

List of Tables

Table 1 – Cumulative Incidence of Recurrence at 5 and 15 Years by Disease and Adjuvant Status.....	18
Table 2 – Meta-Analytic Targets for Calibration of Time-to-Detectable and Clinical Disease.....	32
Table 3 – Estimates of CT Sensitivity by Hepatic Lesion Size.....	41
Table 4 – Distribution of Surgical Outcomes among Patients Undergoing Salvage Surgery.....	51
Table 5 – Cure Rates after Salvage Surgery.....	57
Table 6 – Targets and Model Output for Heterogeneity Parameters.....	61
Table 7 – Salvage Rates and Survival after Recurrence in Unselected Cohort Studies.....	65
Table 8 – Surveillance Schedule for the 4 Arms of the FACS Trial.....	70
Table 9 – Surveillance Schedule for 2 GILDA Trial Arms.....	73
Table 10 – Cohort-Specific All-Cause and Disease-Specific Mortality Hazard Ratios at 5 Years.....	80
Table 11 – Follow-up Regimens under Investigation.....	103
Table 12 – Utilization of Adjuvant and Palliative Chemotherapy by Cohort Age and Stage.....	104
Table 13 – Costs of Surveillance and Treatment of Recurrence.....	114
Table 14 – Optimal Follow up by Cohort and Value of a Life-Year.....	132
Table 15 – Sensitivity Results: Incremental Cost-Effectiveness Ratios.....	138
Table A.1 – Most Recent Professional Society Guidelines for Follow-up.....	167
Table A.2 – Details of Randomized Control Trials.....	169
Table A.3 – Surveillance Schedules of Randomized Control Trials.....	171
Table B.1 – Surveillance Schedule for the Intensive and Minimal Arms Used in Calibrated	192
Table B.2 – Description of Distribution of Sojourn Times by Disease Types.....	193
Table D.1 – Costs of Individual Services.....	197

Table D.2 – Important Utilization Assumptions Regarding Salvage Surgery.....	200
Table D.3 – Aggregate Costs of Surveillance and Salvage and Palliative Treatment.....	202
Table D.4 – Incremental Increases in Life Expectancy (Days) by Cohort.....	205
Table D.5 – Incremental Increases in Mean Total Costs (2017 US \$) by Cohort....	207
Table D.6 – Incremental Cost-Effectiveness Ratios (ICERs) in US \$ per Life Year by Cohort.....	210
Table D.7 – Sensitivity Results: Incremental Cost-Effectiveness Ratios (ICER)....	213

List of Figures

Figure 1 – Schematic Representation of Natural History of Recurrence.....	21
Figure 2 – Sensitivity of a CT Scan and CEA Test vs. Elapsed Sojourn Time.....	43
Figure 3 – Schematic Representation of the Model Implementation of Surveillance...44	44
Figure 4 – Flow Diagram of Treatment of Recurrences.....	48
Figure 5 – Probability Recurrence is Unresectable vs. Elapsed Sojourn Time.....	50
Figure 6 – Cancer-Specific Survival for Unresectable Recurrent Disease.....	57
Figure 7 – Cancer-Specific Survival after Curative Resection by Margin Status and Stage.....	60
Figure B.1 – Forest Plot of the Log Odds of the Chances of Asymptomatic Detection.....	174
Figure B.2 – Forest Plot of the Probability of Asymptomatic Detection in the Intervention Arm.....	176
Figure B.3 – Forest Plot of the Log Odds Ratio of the Chances of R0 Curative Resection among Patients with a Recurrence.....	177
Figure B.4 – Forest Plot of the Probability of R0 Curative Resection in the Intervention Arm.....	178
Figure B.5 – Stage III Colon Cancer No Adjuvant Therapy.....	179
Figure B.6 – Stage III Colon Cancer 5-FU + LV.....	180
Figure B.7 – Stage III Colon Cancer FOLFOX.....	181
Figure B.8 – (Old) Stage II Colon Cancer No Adjuvant.....	182
Figure B.9 – (New) Stage II Colon Cancer No Adjuvant.....	183
Figure B.10 – (Old) Stage II Colon Cancer 5-FU + LV.....	184
Figure B.11– (New) Stage II Colon Cancer 5-FU + LV.....	185
Figure B.12 – Stage II Colon Cancer FOLFOX.....	186
Figure B.13 – (Old) Stage I Colon Cancer.....	187
Figure B.14 – (New) Stage I Colon Cancer.....	188
Figure B.15 – Stage I Rectal Cancer.....	189
Figure B.16 – Stage II Rectal Cancer.....	190

Figure B.17 – Stage III Rectal Cancer.....191
Figure C.1 – Size of Largest Lesion vs. Proportion of Elapsed Sojourn Time.....196
Figure C.2 – Sensitivity of a CT Scan vs. Diameter of Largest Lesion.....195

CHAPTER 1: BACKGROUND & INTRODUCTION

CLINICAL BACKGROUND

Colorectal cancer (CRC) is the third most common form of cancer among US men and women, with about 150,000 new cases diagnosed per year.¹ Almost 80% of cases are confined to the bowel with the possible exceptions of direct extension beyond the bowel wall to adjacent abdominal or pelvic organs (depending upon where the primary tumor is) and/or regional lymph nodes metastases. In the absence of contraindications (poor health due to old age or serious comorbidities), these patients (stage I-III) will undergo curative resection and possibly receive adjuvant chemotherapy or, for rectal patients, neoadjuvant (chemo)radiation therapy. However, depending upon the stage and other factors, between 5%-60% of patients will suffer a recurrence.^{2,3} In the absence of surgical cure, CRC recurrence results in serious morbidity and death.

In colon cancer patients, recurrence generally presents as metastatic disease. The most common site is the liver, but other common sites are the lungs (particularly for distal colon cancer), the peritoneal cavity, or distant lymph nodes.⁴ When colon cancer recurrence is confined to the same quadrant of the abdomen as the primary tumor, it is considered local-regional recurrence and is possibly the result of surgical failure.⁵ Although late-occurring metastatic disease receives most of the focus in the literature, about 10-20% of colon-cancer recurrences represent local-regional recurrence.⁵ Strict local recurrence confined to the bowel, i.e., intraluminal recurrence, used to occur more frequently⁶, but in contemporary practice it is extremely rare (<2% chance).⁷

In rectal cancer, the liver is also the most common site of recurrence, but pulmonary metastases are almost as common.⁴ Although the risk of local-regional failure (in the

pelvis) has decreased dramatically over the last 30 years due to better surgery and neoadjuvant treatment⁸, pelvic recurrence also presents in 15-35% of all rectal cancer patients who suffer recurrence⁹. Like with local-regional recurrence in colon cancer, pelvic recurrence generally presents as primarily extracolonic disease and, though it may involve parts of the lumen, is very rarely confined to remaining portions of the bowel at the time of detection.¹⁰ Pelvic recurrence causes serious morbidity and can quickly become unresectable due to the constricted nature of the pelvis. In particular, the disease may become appended to the sacrum or nervous tissue.¹⁰ In the absence of surgical or, very rarely irradiating, cure, pelvic recurrence will lead to death within 5 years in 90-95% of cases, presumably through metastatic spread to vital organs.^{8 11} Thus from a clinical perspective, unresectable pelvic disease is ultimately just as damming as unresectable distant disease. Though 50-60% of patients who present with pelvic recurrence will also present with (or soon after) distant disease anyway.⁹

Because of the high morbidity and mortality burden associated with recurrence, follow-up surveillance is a common component of modern oncological care. After the primary resection and any (neo)adjuvant therapy for stage I-III colon and rectal cancer patients, routine clinical visits are important for managing treatment-related morbidity and symptoms and providing psychosocial support for patients.¹² However, for the last 40 years, it has also been theorized that early detection of recurrent disease before it is symptomatic can increase the chance of salvage surgery and thereby reduce the morbidity and mortality associated with recurrence. In particular, the hope is to discover the disease while it is isolated in the liver, lungs, or locally and limited enough for salvage surgery with a curative intent.

Currently, the American Society for Clinical Oncology (ASCO)¹³, the American Society of Colorectal Surgeons (ASCRS)¹⁴, the American Cancer Society (ACS)¹⁵, the National Comprehensive Cancer Network (NCCN), the European Society for Medical Oncology (ESMO)¹⁶, and the UK's National Institute for Health and Care Excellence (NICE)¹⁷, among others, recommend at least 3 years, and up to 5 years, of intensive follow-up for patients at high risk for recurrence (stage II and III patients) who are otherwise healthy and would be candidates for salvage surgery. Follow-up efforts are concentrated in the first few years after primary surgery because the risk of recurrence is highest then. Depending on the stage and adjuvant treatment received, 65-95% of recurrences will present within 5 years.^{2 3 18-20} Because the majority (97%-99%) of recurrences are primarily extracolonic in nature²¹, follow-up generally involves surveillance modalities targeting extramural disease: blood tests for tumor markers such as carcinoembryonic antigen (CEA) levels and imaging tests.

CEA is a glycoprotein produced by cells in the gastrointestinal tract which often rises in patients with CRC.¹² CEA levels generally return to normal levels 4-6 weeks after curative surgery, but, in 50-80% of patients with recurrent disease, levels rise prior to the onset of clinical symptoms.^{22 23} Among the patients who exhibit an elevated CEA level prior to onset of clinically recurrent disease, the median lead time of CEA is estimated to be between 3-11 months.^{24 25} CEA rises before symptom onset in most patients with liver disease and retroperitoneal masses (such as in the kidney), but it is less helpful in the case of pelvic and peritoneal disease and a particularly poor indicator for pulmonary recurrence.^{26 27} False positive alerts are also a major problem with CEA levels, particularly among smokers, and so repeated high readings require confirmation with

radiological imaging.²³ Guidelines (Table A.1) recommend CEA monitoring every 3-6 months for at least 3 years and up to 5 years. These are normally performed at routine clinical visits.

To compliment routine CEA testing and clinical exam, guidelines recommend incorporating imaging studies of the liver and lungs (and pelvis for rectal cancer). Abdominal and thoracic (and pelvic) computed tomography (CT) scans are generally used for this purpose, but cheaper options⁶- abdominal ultrasound (US) and chest x-ray - and more expensive options²⁸ – magnetic resonance imaging (MRI), positron emission tomography (PET), or PET/CT combination scan - options are available. Current guidelines (Table A.1) mostly recommend annual thoracic and abdominal (abdominal/pelvic for rectal cancer) CT scans for 3-5 years for stage II and III patients. Thoracic CT was only recently incorporated in the recommendations because of the dissemination and apparent success²⁹ of the practice of pulmonary metastasectomy.

The final component of routine surveillance is endoscopy. In addition to a possible clearing colonoscopy for patients who undergo emergency primary surgery, the US Multi-Society Task Force recommends a colonoscopy for patients of all stages at 1 year and then every 3-5 years if the first colonoscopy is negative.³⁰ The primary motivation for colonoscopy is to detect missed ‘synchronous’ and newly developed ‘metachronous’ high-risk lesions and thereby prevent the development of a second primary CRC. A secondary goal is to detect any recurrence at the site of the surgical anastomosis, though this is extremely rare in current practice.⁷

A REVIEW OF THE CURRENT EVIDENCE

While these recommendations were informed by systematic review and meta-analysis of randomized control trials (RCTs), the strength of evidence was low at best. Moreover, in the last 3 years the evidence has become even more mixed and the optimal surveillance program even more unclear. To date, the results of 16 RCTs comparing alternative surveillance practices have been reported, though only 13 compared more intensive to less intensive surveillance regimens. Two of these 13 trials specifically evaluated alternative endoscopy regimens, i.e., more vs less colonoscopy. The other 11 primarily compared alternative extracolonic-focused regimens. Details of these 13 trials are given in Tables A.2 and A.3. The first trials began enrolling patients in the early-to-mid 1980's³¹⁻³⁴, and the most recent trial's results were published in 2016.³⁵

From the late 1990's through 2015, a series of systematic reviews and meta-analyses found an overall mortality benefit from 'intensive' surveillance compared to 'minimal' or no surveillance.^{6 36-40} The meta-analyses included various subsets of up to eleven RCTs. Most of the trials reported a mortality reduction in the intensive arm, but this result was only statistically significant in two studies^{33 41}. Working with available data, the earlier meta-analyses found an overall mortality reduction at 5 years on average in terms of an odds-ratio, relative-risk, or risk-difference comparing the chances of dying in the intensive arm to the control arm. However, those analyses failed to account for censoring and the relevance of event timing. The last meta-analysis published in that period (2015)³⁷ used established methods^{42 43} to estimate all-cause mortality hazard-ratios (HR), when unreported by authors, using the data presented in the trials. Based on eleven studies, the reviewers found a 25% reduction in mortality hazards associated with intensive surveillance (HR = 0.75; 95% CI: 0.66-0.86). Similar to earlier reviews, they

also found that intensive surveillance decreased the time to diagnosed recurrence, increased the chances of asymptomatic detection of recurrence, and, importantly, increased the chances of salvage surgery of recurrence. This finding was taken to corroborate the theory that increased salvage surgery was driving this remarkably large mortality benefit (very similar in magnitude to that observed for adjuvant chemotherapy).³ Somewhat alarmingly, although there was limited data concerning disease-specific survival, among the trials that did report relevant information, they found no difference between arms.

As a result of these trials and meta-analyses, a near-consensus developed (as evidenced by the professional society guidelines) that intensive CEA follow-up and imaging – modalities directed at detecting pre-clinical extracolonic recurrence – were likely driving the mortality reduction. In spite of the consensus, many questions remained⁴⁴ and the quality of evidence was acknowledged to be suspect⁴⁰. However, this consensus was soon to crumble.

In the last 3 years, two large, recent RCTs - the UK FACS trial^{45 46} and Italian GIDLA trial³⁵ - that evaluated intensive follow-up regimens with particularly intensive imaging reported no overall survival benefit in the intensive arm. The FACS trial used a 2x2 factorial design to compare 5 years of biannual or annual CT scans vs. just 1 CT scan at 12-18 months and intensive CEA vs. no CEA follow-up. There were thus 4 different arms: intensive CT and CEA, intensive CT alone, intensive CEA + 1 CT scan, and minimal surveillance (1 CT scan). They found that the 3 more intensive arms (CT & CEA, CT Only, and CEA + 1 CT) were more likely to detect recurrence while asymptomatic and more likely to perform salvage surgery ‘with a curative intent’ than the

minimal surveillance arm (i.e., 1 CT scan only). However, somewhat surprisingly, there was a small, non-significant overall and cancer-specific survival advantage for the minimal surveillance arm.

The GILDA trial compared more frequent to less frequent abdominal ultrasound, colonoscopy, chest x-ray, and, for rectal patients, abdominal/pelvic CT. Both arms were subjected to intensive CEA testing. The more intensive arm detected disease earlier on average, but it did not significantly increase the rate of asymptomatic detection or salvage surgery. Cancer-specific survival is not reported, but again there was a slight, non-significant mortality advantage for the less-intensive arm.

Moreover, the results of an old, previously unpublished and prematurely terminated RCT from the 1980's and early 1990's – the CEA Second Look (CEASL) trial – were recently published. The CEASL trial enrolled roughly 1,500 patients with routine CEA assessment. A little over 200 patients with repeated elevated CEA measurements were randomized to a conservative strategy that waited for symptoms to develop or an aggressive strategy that, after poorly-defined clinical assessments, performed exploratory laparotomy ('second-look surgery'). This trial reported no survival benefit from CEA-driven second-look surgery among those randomized to aggressive treatment.

After incorporating these recent trial results and making alternative exclusion decisions (different than earlier reviewers), two^{47 48} subsequent meta-analyses have concluded that intensive surveillance does not provide an all-cause or disease-specific mortality benefit. The disappointing results of these trials and the new meta-analytic perspectives have unsurprisingly lead for calls to throw in the towel on intensive follow-

up. Discussing the professional-society recommendations (Table A.1), the authors of the report on the GILDA trial conclude that

“Results from GILDA, FACS and CEA Second-Look randomized trials failed to support such recommendations and they undermine the paradigm that earlier detection of recurrences may translate into either longer survival duration or enhanced quality of life in patients with colorectal cancer treated with curative intent”³⁵

The authors of the two latest meta-analyses dismiss the logic behind intensive surveillance and even express skepticism towards any role for follow-up with the goal of detecting recurrences earlier:

“The information we have now suggests that there is little benefit from intensifying follow-up... We do not know what is the best way to follow patients treated for non-metastatic colorectal cancer, or if we should at all.”⁴⁷

“Based on pooled data from randomized trials published from 1995 to 2016, the anticipated survival benefit from surgical treatment resulting from earlier detection of metastases has not been achieved.”⁴⁸

Critics have theorized that the reason (intensive) surveillance does not provide a survival benefit is because metastasectomy is unlikely to improve survival, at least in most cases.⁴⁹ They dismiss the many surgical case-series reporting favorable outcomes among patients receiving hepatic and pulmonary resections as the result of extreme patient selection and suggest that many of these patients would have been long-term survivors in the absence of resection due to a slow-moving disease natural history.⁵⁰⁻⁵²

Another consistent and less extreme explanation would be that hepatic, pulmonary, and local resection do benefit a limited number of patients but that these patients have slow moving disease that, in general, would have been similarly salvageable if detected later by less intensive surveillance. Moreover, the extra-intensive surveillance regimens

investigated in these necessarily non-blinded trials had the result of lowering the bar for what is considered salvageable disease. Surgeons detect disease earlier than normal and so, as believers in their own craft, more aggressively attempt salvage surgery. Thus salvage rates go up. However, at the margin, these additional salvages are not helpful because these patients have a disease with a more aggressive biology or that is already more widely disseminated. Thus, at the margins, there is no survival benefit and potentially a limited harm due to surgical-related morbidity and mortality.

The FACS trial investigators hypothesized that much of the overall mortality benefit picked up in earlier meta-analyses could be due to early detection of patients with residual macro-metastases or local disease after surgery for the primary CRC tumor.⁴⁶ That is, they speculated that intensive surveillance was helpful when it picked up missed synchronous metastatic disease (stage IV) but less so when it picked up newly-appearing metachronous metastatic disease. This is relevant because the accuracy of diagnostic imaging and staging has improved over the last 30 years and both the GILDA and FACS trials enrolled patients after any adjuvant treatment (up to 6 months after primary surgery) and used extensive diagnostic imaging to rule out missed systemic disease. However, it's difficult to reconcile this hypothesis with the observations that metachronous metastases fare better than synchronous metastases⁵³ and that, even among metachronous cancers, for the first 3 years, time-to-recurrence is inversely correlated with survival after recurrence.⁵⁴⁻⁵⁶

Whatever the explanation, critics have pointed to the negative results of the GILDA and FACS trials and reasonably argued that, if we do not see any benefit in these trials, we are unlikely to find a benefit in future research. This is because these trials (a) are by

far the largest, (b) evaluate the most intensive imaging regimens, and (c) are the most recent trials and thus best reflect the modern surgical capabilities for salvage and the modern disease process. This last component is important because the disease has changed dramatically over the last 40 years with the adoption of modern (neo)adjuvant therapies and improved primary surgery.

It is likely that future iterations of professional guidelines will incorporate these results and conclusions into their recommendations, and this in principle could lead to a large-scale change in oncological practice in North America and Europe. Because of the large decline in recurrence rates, the disease-specific excess mortality hazard faced by non-metastatic CRC patients has declined notably over the last 35 years for all stages and for both colon and particularly rectal cancers.⁵⁷ This is likely due to better surgery, e.g., Total Mesorectal Excision⁹, (neo)adjuvant therapy¹⁸, and stage-migration⁵⁸. One important consequence of this heartening development has been an increasing number of survivors of stage I-III colorectal cancer. A recent estimate puts the current number of CRC survivors in the US at almost 1.5 million.⁵⁹ Thus any change in surveillance recommendations could have large-scale implications on practice patterns and resource utilization. It is therefore paramount that available evidence, as problematic as it is, is interpreted appropriately. The stakes are particularly high for younger CRC patients (<55) - a group on the rise (particularly for rectal cancer)⁶⁰ – who, in the absence of CRC, could expect to live multiple additional decades.

REEXAMINING THE LOGIC OF FOLLOW-UP

In this dissertation, I argue that the current discussion of available evidence surrounding intensive surveillance suffers from a failure to carefully work through the

implications of the logic underpinning CRC follow-up. As mentioned above, intensive follow-up has been theorized to lead to earlier detection of recurrent disease that is more likely to be isolated and salvageable. However, it is unclear that this theory is consistent with the sizable all-cause mortality benefit that was reported in older meta-analyses and that recent trials were powered to detect. A simple back of the envelope calculation highlights the problem. Assuming a 5-year cumulative incidence of recurrence of roughly 30% among stage III colon cancer patients (treated with FOLFOX adjuvant chemotherapy)¹⁹, and assuming that 5-10% of recurrences (almost all local-regional) will be salvageable in the absence of follow-up (when clinically-indicated)^{5 11 61}, then, if an intensive follow-up regimen was able to achieve macroscopically and microscopically clear margins (R0 resection) in 50% of patients with recurrence, at most 12-14% of a cohort could benefit from intensive follow-up.¹ If we consider a more realistic situation - intensive versus less intensive follow-up, a mix of stage I-III patients (20% recurrence), and a more realistic R0 salvage rate of around 40% in the intensive arm (vs. 15% in minimal arm) – we might expect only about 5% of patients to benefit.²

This simple thought experiment suggests an alternative explanation for the negative results of the two recent trials. Perhaps the absence of any significant mortality benefit in the intensive follow-up arms of recent trials is not due to the failure of salvage surgery (resulting from earlier, asymptomatic detection of recurrence) to confer any benefit to such patients. Rather, perhaps it is due to the small proportion of patients who can possibly benefit from such treatment.

¹ $0.3 * (0.5 - 0.1) = 0.12$ & $0.3 * (0.5 - 0.05) = 0.14$

² $0.2 * (0.4 - 0.15) = 0.05$

In Chapter 3 of this dissertation, I argue for this latter explanation. I first argue there is reason to believe in the efficacy of salvage surgery for isolated recurrent disease, even when performed at the aggressive levels seen in the FACS and GILDA trials. Moreover, I use a modeling analysis to show that the failure of these trials to detect any significant mortality reduction is not surprising and should not be interpreted as incompatible with the hypothesized efficacy of follow-up-induced salvage surgery. In particular, I estimate the true incremental mortality benefit we could expect in moving from the FACS minimal follow-up arm to the more intensive arms. I perform a similar analysis in the GILDA trial. I also highlight and evaluate a potential source of bias in the FACS trial that would likely have favored the minimal follow-up arm: a recurrence-imbalance. In light of this bias and the low recurrence rates, I argue that the trials were severely underpowered to detect the small hypothesized true benefit.

The analyses of Chapters 3 and 4 rely on a microsimulation model of CRC recurrence, detection, treatment, and mortality. The model has been constructed to embody the hypothesis that more aggressive extra-colonic follow-up of patients can increase the proportion of recurrences that are amenable to curative resection and thereby improve the survival of such patients. However, it is also empirically grounded in the sense that (a) it can replicate many important features of the natural history process of recurrence, (b) the increase in salvage rates attributable to intensive follow-up in the model are consistent with meta-analytic results, and (c) the survival experience of patients treated with salvage surgery is consistent with that observed in unselected cohort studies and, importantly, the FACS trial.

If I am right that, while intensive-surveillance-induced salvage surgery can benefit select patients, in current practice the ex-ante expected benefit of (intensive) surveillance for any given patient will be too small to detect in any realistically-powered RCT, the practice is likely to be dismissed as not clinically relevant or at least unlikely to be cost-effective. However, the same is true in many cases of cancer screening.³ For example, a trial with more than 68,000 participants failed to find an overall-mortality reduction from fecal occult blood testing⁶² yet it is considered a cheap and effective screening method in some populations.⁶³ When a treatment benefit is (very) small, whether or not to use it and at what dosage are questions that require a decision-analysis that considers both the incremental benefits and incremental costs of varying intensities of the intervention and incorporates our uncertainty about these benefits and costs into the analyses. *A priori*, it does seem very unlikely that the level of intensive imaging tested in the CT arms of the FACS trial (7 CT scans in 5 years) would be cost-effective given the likely diminishing returns associated with each additional CT scan. However, that does not preclude the possibility that the expected benefit procured by one, two, or even three CT scans (over and above routine CEA testing) might be considered worth the cost. Either way, a careful decision analysis is needed.

In Chapter 4 of this dissertation, I use a modeling analysis to conduct a cost-effectiveness analysis of several follow-up strategies. In particular, I consider (i) no follow-up, (ii), a single CT scan at 12-18 months, (iii) guideline-level CEA testing and a single CT scan at 12-18 months, and guideline-level CEA testing and annual CT scans

³ This point was suggested to me by Professor Karen Kuntz.

for (iv) 2, (v) 3, and (vi) 5 years. I conduct these analyses separately for stage II and III patients and colon and rectal patients. Analyses use a lifetime time horizon and are undertaken from the perspective of the US healthcare system.

The remainder of the Dissertation is organized as follows. In the next chapter I describe the model that was developed for this dissertation. This includes a discussion of the model structure, data sources, and calibration efforts and results. Chapters 3 and 4 assume the reader has read the above chapter as well as Chapter 2.

CHAPTER 2: THE MICROSIMULATION MODEL

OVERVIEW OF THE MODEL

I developed a discrete-event microsimulation model of CRC recurrence, recurrence-detection, treatment, and mortality. The model was constructed and, when necessary, calibrated in the open source R software.⁶⁴ The model simulates the life-course of colorectal cancer patients who have successfully underwent and survived curative resection of the primary tumor. The main processes of the model involve (i) the natural history of extracolonic recurrence, (ii) surveillance testing and detection of recurrence, (iii) treatment of recurrence, and (iv) death from cancer or other causes based upon the treatment received. The natural history process of recurrence is modeled separately for rectal and colon cancer patients, for each stage (I, II, and III), and, in the case of colon cancer, by adjuvant treatment received.⁴ Unless otherwise stated, the other processes are identical for rectal and colon cancers, for each stage, and regardless of any adjuvant therapy received.

The model focuses on extramural disease and does not currently include intraluminal, e.g., anastomotic, recurrence or surveillance (endoscopy). With current surgical practice, isolated intraluminal recurrences are very rare. For example, in the FACS trial, less than 2% of recurrences were detected by endoscopy. Endoscopic surveillance is generally

⁴ As will be described below, more and better targets were available for colon cancer recurrence. For stage II-III rectal cancer, the targets represent the risk of recurrence for patients undergoing neoadjuvant and/or adjuvant treatment. For stage II-III colon cancer, targets were available for patients treated with surgery alone as well as for patients treated with adjuvant therapy.

justified for the purposes of detecting and preventing second primary lesions and thus represents an extension of the logic of CRC screening.³⁰

The model also does not distinguish between local-regional (extra-colonic) disease and metastatic disease. In the case of rectal cancer, this choice might be questioned. However, for the purposes of this dissertation, the two types of disease are sufficiently similar to be represented as one process. In particular, both forms of disease are fatal in the absence of surgical cure, both forms of disease are routinely detected by CEA measurement and imaging studies, a roughly similar proportion of cases are amenable to curative resection, and the survival experience of patients after curative resection is comparable.^{11 45 61 65-67} However, due to the higher incidence of local-regional disease in rectal cancer patients, the hazard of any recurrence is higher in stage I and II rectal cancer patients than in comparable colon cancer patients. The model incorporates this difference by modeling colon and rectal cancer patients separately.

A microsimulation approach was selected over a Markov cohort model because of the importance of incorporating heterogeneity in several component disease processes and correlation among those processes. For example, from the perspective of evaluating surveillance testing, an important phenomenon of the natural history process of recurrence is that patients who develop clinical disease earlier face a higher mortality hazard than do patients who develop recurrence later.⁵⁶ A discrete-event framework was chosen over the classic state-transition approach because of its speed and flexibility.⁶⁸ A discrete-event model is organized around events rather than disease-states and transitions between those states.⁶⁹

In what follows, I describe each of the main model components mentioned above: (i) recurrence natural history, (ii) surveillance testing and detection of disease, (iii) treatment of recurrence, and (iv) death from cancer depending upon the treatment received. In each section, I describe key phenomenon the model is meant to capture, how this is implemented in the model, and parameters governing the process. When possible, parameters were taken from the literature. However, in some cases, I used calibration, i.e., ‘reverse estimation’ to derive parameter values that lead to model output that best fit data targets in some sense. The logic here is that some parameters are not directly estimable for ethical or logistical reasons. For example, parameters governing the distribution of sojourn times for a cancer (the period during which the disease is potentially detectable but pre-clinical) cannot generally be directly estimated because once a disease is detected it will inevitably be treated. In this case, a calibration approach uses indirect evidence to constrain such parameters, and thus the model, to at least be consistent with (able to recreate) observable data.

THE NATURAL HISTORY OF RECURRENCE

Even in the case of stage 3, most patients who receive modern (neo)adjuvant therapy regimens will not develop a recurrence. Analogous to the principles of cure models that have been used recently for survival analysis⁷⁰, the microsimulation model initially divides patients into ‘cured’ and ‘not cured’. The latter patients are disease free for the entirety of the model simulation and will die of other causes. The non-cured patients will develop clinical recurrent disease in the absence of surveillance within the next 15 years. Although most recurrences occur within 5 years, late-occurring recurrence (after 5 years and up to as late as 15 years) has been documented as more common than previously

thought.^{2 3 18 71} The chances of suffering a recurrence over the next 15 years vary by stage, location (rectal vs. colon), adjuvant treatment, and (as explained below) period (old vs. newer). These parameters are presented in Table 1 along with recurrence rates at 5-years for comparison. Both the 5-year and 15-year recurrence rates were taken from the recurrence-targets discussed below. In the case of survival curves, these values were readily available. In the case of time-varying hazard functions, I used numerical integration to estimate them.

Table 1: Cumulative Incidence of Recurrence at 5 and 15 Years by Disease and Adjuvant Status

Disease	Adjuvant Therapy	Proportion Recurring within 5 Years	Proportion Recurring within 15 Years	Sources
Stage III Colon	No Adjuvant	0.555	0.597	3 18
Stage III Colon	5-FU + LV	0.366	0.412	3 18 19
Stage III Colon	FOLFOX	0.298	0.349	3 18 19
Stage II Colon (New)	No Adjuvant	0.180	0.210	3 18 19 58
Stage II Colon (Old)	No Adjuvant	0.225	0.267	3 18
Stage II Colon (New)	5-FU + LV	0.144	0.176	3 18 19
Stage II Colon (Old)	5-FU + LV	0.197	0.228	3 18 58
Stage II Colon	FOLFOX	0.097	0.130	3 18 19
Stage I Colon (New)	-	0.056	0.085	2 3 18 20
Stage I Colon (Old)	-	0.077	0.114	2 3 18 20
Stage III Rectal	Neoadjuvant Chemoradiation Therapy and Adjuvant Chemotherapy Therapy with FOLFOX	0.289	0.320	9 72

Stage II Rectal	Adjuvant 5-FU + LV & Possibly Neoadjuvant Radiation Therapy	0.204	0.232	9 73
Stage I Rectal	---	0.083	0.115	20 74

BACKGROUND THEORY

Most patients who develop recurrence after primary surgery present with metachronous metastatic disease. By ‘metachronous’ I mean to refer to late-occurring metastatic disease as opposed to synchronous (same-time occurring) metastatic disease, i.e., stage IV CRC. The current theory is that these patients have dormant micro-metastases implanted in some part of the body (in an organ, tissue, cavity, or lymph nodes) or at least circulating in the blood prior to the detection and surgical resection of their primary CRC. While these cancer cells likely have already infiltrated the organ, they lie dormant for some latency period because the micro-environmental and genetic conditions of the new location are not conducive for lesion growth. In the cases where the micro-metastases have no or only a very short latency period, the disease develops quickly and presents as synchronous metastatic disease. Initially, these micro-metastases (possibly just a few malignant cells) are undetectable by surveillance tests (e.g., CT imaging) and thus are referred to as ‘occult’. However, at some point the micro-environmental and genetic conditions will align for the disease to begin to colonize the organ and transition into a pre-clinical and then clinical macro-metastasis.⁷⁵

A small percentage of recurrences (particularly for stage III patients) will actually represent missed synchronous metastases, i.e., misdiagnosed stage IV disease. That is,

these patients will have a macro-metastasis in say the liver which was missed by diagnostic imaging (imperfect sensitivity) done prior to surgery (or after emergency surgery). These patients will generally present with clinical disease within 6 months.⁴⁶ The remaining patients who suffer a recurrence (other than metachronous or missed synchronous metastatic disease) will present with local-regional disease. This is located in the pelvis for rectal cancer and the abdomen for colon cancer, and it is thought to be a result of surgical failure or failure to detect regional lymph node metastases and treat accordingly.^{5 67}

The model explicitly focuses on the development of metachronous metastases. There are three key processes after the initial implantation of the micro-metastasis in a distant organ: (a) initiation of growth, e.g., the cell division rate overtakes the cell death rate, (b) the continuous growth process, and (c) the onset of clinical disease. For the purposes of a model evaluating surveillance testing, it is useful to represent the above set of processes as two periods (Figure 1): (1) the latency period - the time during which the (possibly dormant) micro-metastasis is undetectable by surveillance modalities - and (2) the sojourn period – the time during which the macro-metastasis is growing, is detectable by surveillance tests, yet is pre-clinical, i.e., asymptomatic. The first period is demarcated by the beginning of the simulation model and the simulated time at which the metastasis becomes ‘detectable’. The second period is demarcated by the simulated time between which the disease becomes detectable and the disease becomes ‘clinical’. Here we assume that the patient suffers symptoms and seeks care and/or would show signs of recurrence upon clinical examination. The process of lesion growth and progression is

relevant for detection by surveillance and the prospects of salvage surgery, respectively, and so is covered in a latter section.

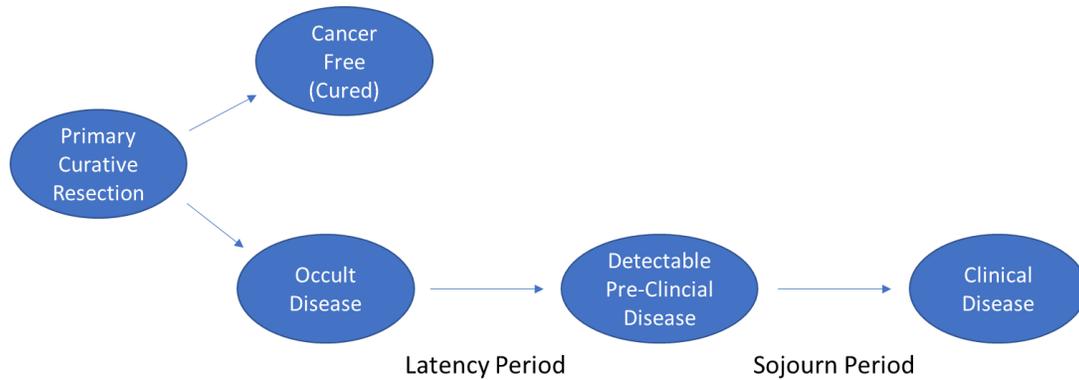


Figure 1: Schematic Representation of Natural History of Recurrence

MODEL STRUCTURE & PARAMETERS

I note that, although the model is meant to represent metachronous metastases, the above representation should be sufficiently abstract to adequately capture the development of local-regional recurrence. Moreover, the model implicitly accommodates missed synchronous metastatic disease by allowing significant heterogeneity in the time-to-detectable and time-to-clinical disease processes and correlation among those processes. Thus, some individuals in the model will develop detectable disease immediately (within less than a month), and these individuals will also tend to have a shorter sojourn time. More specifically, the model includes three heterogeneity terms (i.e., random effects). One for the time-to-detectable-disease process (duration of latency period) and one for the time-to-clinical-disease process (duration of sojourn period). A third heterogeneity term is also included for time-to-death processes and is described below. The heterogeneity terms are assumed to modify the respective baseline hazard rates multiplicatively (additively on the log scale). They are simulated on the log-scale

from a mean-zero multivariate normal distribution with a single shared correlation parameter.

The key processes to simulate in the natural history model of recurrence are thus the time to detectable disease and the time to clinical disease (conditional on the disease being detectable) among those who will eventually suffer a recurrence. I chose to directly model the hazard rates for each process because the best available targets (indirect evidence) were generally hazard rates of recurrent disease. These targets are discussed below. One disadvantage of using a discrete-event simulation framework is that a modeler is generally forced to rely on parametric models for ease of simulation. However, in the absence of direct evidence on the time-to-detectable and time-to-clinical disease processes (among those suffering a recurrence), it is difficult to anticipate what parametric survival models might be appropriate.

The microsimulation model avoids the problem of selecting a parametric model by relying on nonparametric methods to model the hazard rates for the development of detectable and clinical disease among those who will ultimately develop a recurrence. In particular, I used cubic B-splines as they have been used successfully in the survival literature to model hazard functions.⁷⁶ Splines are piecewise polynomials of degree k that are continuous at intersecting points ('knots') in the sense that derivatives are continuous up to degree $k-1$. Cubic B-splines are convenient and efficient basis functions with which any spline function of degree 3 (for a given set of knots) can be uniquely represented.⁷⁷ Therefore, a wide range of functions can be efficiently approximated by a linear combination of cubic B-splines. For a given set of points in the function domain (a grid of times in the domain of the hazard function in our case), set of knots (breakpoints), and

function degree (cubic in our case), a basis matrix can be generated using R's spline package. The key parameters governing the hazard rate models are thus a set of linear coefficients, and so parameterizing the model requires calibrating these.

In order for a non-parametric model of the hazard function to be useful in a discrete-event context, it must be possible to simulate from the implied distribution. A convenient feature of using cubic B-splines is that an analytic solution exists (and is available in R) for definite integrals of the hazard rate. The definite integrals of the hazard function represent the cumulative hazard and can easily be converted to a survival function. Thus the key modeling task is to simulate from the survival function. In early modeling efforts, I relied on rejection sampling⁷⁸ to sample from the implied density function (equal to the product of the hazard and survival functions). However, because of the unwieldy form of calibrated density functions, this method was very inefficient. I instead opted for a discretization approach in which event times were simulated to fall within roughly 3-day intervals (one tenth of a month). A comparison of methods suggested this resulted in only trivial bias.

CALIBRATION TARGETS AND METHODS

The main obstacle to parameterizing the time-to-detectable- and clinical-disease processes (among those who would eventually suffer a recurrence) was the lack of any direct evidence on the distributions characterizing these two processes. Available targets provide an estimate of the hazard rate at which patients present with recurrent disease, i.e., the *time-to-diagnosed-recurrence process*. However, they do not provide information about when a patient's disease became detectable and thus the duration of each component process. If we condition simply on the *time-to-diagnosed-recurrence* data

targets, in addition to being inestimable, the parameters governing these two component processes are obviously non-identifiable.⁷⁹ In a calibration setting, this means that multiple different parameter sets will be able to reproduce target data equally well and thus it is unclear which set to prefer. The constraint of having to be able to reproduce additional data targets – meta-analytic rates of asymptomatic detection and salvage surgery from the surveillance trials – can be expected to constrain the feasible parameter space, but the parameters are still likely to remain nonidentifiable.

When confronted with nonidentifiable parameters in need of calibration, several options are available. One is to place constraints on the parameter space, ideally guided by theory or content knowledge so that they become identified. Another option is to use informative priors and utilize Bayesian calibration methods.⁷⁹ However, in our case, these would likely be arbitrary and poorly informed. Another option is to calibrate one acceptable parameter set and simply use that. A weakness of this approach is that different sets of equally well-fitting calibrated parameters might lead to different modeling conclusions. A superior approach follows the principles of Bayesian model averaging⁸⁰ and attempts to average over this uncertainty. This is the approach I selected to calibrate the natural-history model of recurrence. In what follows, I first describe the data targets and then provide details of the calibration process.

The linear coefficients governing the time-to-detectable- and time-to-clinical-disease hazards among those who will ultimately suffer a recurrence were calibrated to two data targets: (i) continuous-time hazard functions or survival curves of recurrence taken from RCTs of (neo)adjuvant treatment and (ii) meta-analytic rates and odds-ratios (comparing treatment arms) of asymptomatic detection of recurrence and salvage surgery from the

intensive surveillance trials. In some cases, the *time-to-diagnosed-recurrence* targets represented the incidence of disease in the absence of routine follow-up, and in other cases the trial protocols called for routine surveillance. In the latter case, calibration was performed with the same follow-up structure in the model as was described in the target source. The surveillance schedules used for the purposes of calibration are described in the captions under the figures depicting the *time-to-diagnosed-recurrence* targets (Appendix B). These are described in the results section below. Forest plots of the meta-analytic targets are given in Figures B.1-B.4. They were based upon my own meta-analytic analyses of trials that evaluated intensive extracolonic surveillance strategies. The purpose of these meta-analytic targets is to constrain the distributions of the two component processes (time-to-detectable- and clinical-disease) such that realistic levels of surveillance-based detection and salvage of recurrence can occur. I chose not to use the meta-analytic all-cause mortality hazard-ratio as a target because, conditional on the rate of surveillance-based detection and salvage targets, cancer-related mortality is largely determined by processes governed by other parameters and thus is only tangentially related to the underlying development of recurrence.

For colon cancer, more detailed and precise time-to-recurrence targets were available. Stage II and III colon targets were derived from publications by the Adjuvant Colon Cancer End Points (ACCENT) Collaborative Group.⁸¹ The ACCENT database was first conceived in the 1990's by researchers at Mayo Clinic and has grown to contain individual-patient level data for over 40,000 patients from 27 RCTs of adjuvant chemotherapy for colon cancer conducted between 1977 and 2009. Using this database, researchers have published multiple analyses of aspects of adjuvant chemotherapy

treatment pooling individuals from across all relevant trials. I took *time-to-diagnosed-recurrence* targets from several such publications. In particular, I used the open-source software PlotDigitizer to digitize continuous-time hazard functions or, if not available, survival curves for recurrence and then implemented them in R using spline interpolation. In most cases multiple sources were used to construct a target hazard function or survival curve that extended to 15 years.

For stage III colon cancer, continuous-time hazard-function targets were available for patients treated with surgery only, Fluorouracil and Leucovorin (5-FU/LV), and FOLFOX (5-FU/LV + Oxaliplatin). The target hazard function for stage III colon cancer patients treated with surgery alone came from a pooled meta-analysis of Fluorouracil-based regimens in an adjuvant setting.³ In particular, years 1-8 of the target were backed out from the combination of a disease-free-survival (DFS) hazard function for stage III patients ($N > 1,600$), a recurrence hazard function combining both stage II and stage III patients ($N = 2,500$), and information about stage-specific background mortality and recurrences. These patients were enrolled in trials in the 1970's and 1980's and were assumed to undergo no routine surveillance testing. The full constructed target is depicted in Figure B.5. Years 9-15 were backed out from a pooled hazard function combining stage II and III patients³ and information concerning the relative risk of late-occurring recurrence.¹⁸ This latter component was not specific to any adjuvant treatment as there is no evidence of any benefit from adjuvant therapy with regards to late-occurring recurrence.^{3 18 19} It was therefore used (often in a slightly adjusted form) for all other disease targets. For stage III patients treated with 5-FU/LV and FOLFOX, years 1-5 of the target hazard functions were taken from a patient-level meta-analysis¹⁹ of five RCTs

of FOLFOX with $N > 2,400$ and $N > 7,500$ stage III patients, respectively. Years 5-8 were taken from the same source as was used for surgery alone. The corresponding targets are depicted in Figures B.6 and B.7, respectively.

For stage II colon cancer, hazard functions (or survival curves) were available for patients treated with surgery alone and again for patients treated with 5-FU/LV and FOLFOX. Moreover, there was evidence of a reduction in the hazard of recurrence in newer trials that was unrelated to treatment (both among patients treated with surgery alone and those treated with 5-FU/LV).⁵⁸ This was likely the result of stage migration due to improved diagnosis of regional lymph node metastases, and so I calibrated both a new and old set of hazard functions. The old target for stage II colon cancer treated with surgery alone came from the same sources ($N > 800$) as described for stage III colon cancer with no adjuvant and is depicted in Figure B.8. The first 2.5 years of the corresponding new target (Figure B.9) came from the same source³ but was adjusted by a HR (representing the effect of stage-migration) from a second source⁵⁸ using the ACCENT database. The remainder of the target replicated the equivalent portion of the new target for stage II colon cancer treated with 5-FU/LV (described below) as there is no evidence of a chemotherapeutic-induced recurrence reduction after 2-3 years for stage II patients.³ Years 1-6 of the (new) target hazard functions for stage II patients treated with 5-FU/LV (Figure B.11) and FOLFOX (Figure B.12) were taken from the previously-mentioned patient-level meta-analysis³ of 5 RCTs (ACCENT database) investigating FOLFOX in an adjuvant setting. The targets represented the disease experience of $N = 800$ and $N = 1,450$ stage II patients, respectively. Years 6-8 were taken from the previously described meta-analysis of 5-FU-based regimens.³ Finally, Figure

B.10 depicts the survival curve target characterizing the (old) recurrence experience of stage II patients treated with 5-FU/LV.⁵⁸

It was necessary to calibrate different sets of hazard functions (time-to-detectable and clinical-disease) for different adjuvant therapies (even within the same stage) because the benefit of adjuvant therapy in terms of avoided recurrences appears to fall almost exclusively within the first 2 years after primary surgery.³ A simple adjustment to the model's cure rate or an adjustment of the hazard function with a constant hazard ratio (HR) would therefore misrepresent the recurrence hazard under adjuvant therapies and potentially limit the validity of subsequent modeling analyses comparing alternative surveillance schedules.

For stage I colon and rectal cancer, a time-to-recurrence survival function was taken from a large hospital database.²⁰ For colon cancer, there was again evidence of a reduction in the risk of recurrence over time, and so I calibrated a new and old set of hazard functions.² The targets are shown in Figures B.13-B.15. For locally-advanced rectal cancer (stage II^{9 73} and III⁷²), targets represented the disease experience of patients treated with Total Mesorectal Excision surgery. The target for stage II rectal cancer (Figure B.16) was derived from a RCT evaluating 5-FU/LV in mostly stage II rectal patients (N > 1,450). Patients were treated with adjuvant 5-FU/LV and possibly neoadjuvant radiation therapy. The target for stage III rectal cancer (Figure B.17) came from a RCT evaluating FOLFOX and represented the disease experience of patients (N > 450) treated with both neoadjuvant chemoradiation therapy and adjuvant chemotherapy using FOLFOX. Separate targets for different combinations of neoadjuvant and adjuvant therapy were unfortunately not available.

The basic outline of my approach to calibration for each disease location (colon vs. rectal), stage, and adjuvant therapy (when available) combination was as follows. I firstly identified a set of pairs of vectors of linear coefficients that together could reproduce the appropriate *time-to-recurrent-disease* target. The goal was to select a set of pairs of time-to-detectable- and time-to-clinical-disease hazard functions (among those getting a recurrence) that covered the feasible range of latency times and sojourn times while still matching the target. The next step was to assign a discrete probability distribution to these pairs of parameters. To do this, each such pair was weighted by the product of a measure of its degree of fit with the *time-to-recurrent-disease* target and a measure of its compatibility with meta-analytic targets. Finally, the set of weights were normalized to sum to 1.

In what follows I provide a more detailed sketch of the calibration process. I evaluated pairs of vectors of linear-coefficients (one for each disease process) by quantifying the discrepancy between the model-output they produced and the respective targets. In order to identify a wide range of pairs of plausible time-to-detectable- and time-to-clinical-disease hazard functions, I firstly calibrated a baseline pair assuming they were equal. That is, I applied an equality constraint to coefficients for the detectable and clinical disease processes so that the hazard functions of both processes would be identical. I then used a variant of simulated annealing available in R's `optim` routine to find the set of coefficients and the two relevant heterogeneity variance parameters⁵ that minimized the mean absolute

⁵ Early exploratory analyses suggested that a very strong positive correlation among heterogeneity terms would need to be assumed to match targets (described below) concerning the relationship between time to

difference between the model-produced *time-to-diagnosed-recurrence* hazard function (or survival function) and the data target, evaluated at a grid of time-points. Simulated annealing is a stochastic optimization approach that can provide good approximations to a global optimum.⁸² Exploratory analyses suggested a Monte Carlo size of 300,000 per model run was sufficient to achieve stable results. In the case of a hazard-function target, model-produced times of diagnosed recurrence were smoothed using nonparametric methods (R's *bshazard* package) just as they were in the target sources. This smoothed hazard function was then compared to the target at a fine grid of time points. In the case of a target (recurrence-free) survival function, I used model-output to construct a Kaplan-Meier (KM) survival curve which was then compared with the target curve over a grid of points.

This first step provided one good set of parameters: both heterogeneity variances and linear coefficients for the baseline hazard functions. However, as stated above, the goal of the calibration exercise was to cover a sufficiently wide range of good parameter sets so that the non-identifiability problem could essentially be averaged over. For convenience, I used the heterogeneity variances calibrated in the first step for all additional parameter sets. This was deemed inconsequential because, over a relatively wide range of values, variation in these parameters had very little impact on model fit with respect to *time-to-recurrent-disease* targets. Moreover, their value was later adjusted to fit a different target (discussed below). Thus for the second step, I semi-formally adjusted the calibrated linear coefficients in opposite directions to construct parameter sets which lead to longer latency periods and

recurrence and time to death. I thus calibrated the heterogeneity variances for detectable/clinical disease using a working correlation of 0.8

shorter sojourn times as well as parameter sets which lead to longer sojourn times and shorter latency periods. In general I did this by multiplying the baseline pair of coefficients by incrementally larger (smaller) factors (increments of +/- 5%) until there was serious deterioration in fit as judged subjectively.⁶ Occasionally some manual tuning and/or formal calibration of one or more coefficients was needed to preserve reasonable fit. In particular, when sojourn times were increased significantly on average (meaning the disease spent longer developing into clinical disease), relatively much larger decreases in latency times on average were required to preserve fit (meaning the speed with which occult disease grew into potentially detectable macro-metastases had to be greatly accelerated).

The final step was to assign these (generally 15-25) parameter sets a discrete probability mass function. For this purpose, I used a product of (a) the goodness-of-fit of the model-produced hazard function (or survival curve) with the data target (mean absolute difference across a grid of points) and (b) a multivariate likelihood value based upon fit to meta-analytic targets. For both (a) and (b), values were normalized to sum to 1. Finally, the resulting product of the two normalized values was itself normalized to sum to 1. At this point, any set of coefficients with a probability mass of less than 0.01 was excluded and weights were renormalized.

For component (b), I used meta-analytic predictive distributions to evaluate the model-produced (1) proportion of recurrences that could be (R0) salvaged in an intensive-surveillance arm representative of the admittedly heterogeneous set of

⁶ While the reliance on subjective judgment might be concerning to some, these judgments were eventually supplanted by a formal rule that excluded parameter sets with an assigned probability of less than 0.01. This is discussed below.

interventions investigated in the RCTs and (2) odds ratio comparing the intensive arm to a representative minimal arm. The respective surveillance schedules used are depicted in Table B.1. Similar likelihood values (based on the meta-analytic predictive distribution) were assigned to the model-produced (3) proportion of recurrences detected asymptotically, i.e., by extracolonic-focused surveillance (CEA assays and CT tests), in the intensive arm and (4) the analogous odds ratio.⁷ A meta-analytic predictive distribution combines our uncertainty of the location of the population-average effect with the heterogeneity estimated to exist in the population (and, by using a T-distribution, our uncertainty surrounding our estimate of the heterogeneity).⁸³ It thus represents the range of possible true parameter values we would expect to see in the real world. Table 2 gives the relevant meta-analytic population-average point estimates, 95% confidence intervals for the population-average effect, and 95% prediction intervals using a random-effects model.

Table 2: Meta-Analytic Targets for Calibration of Time-to-Detectable- and -Clinical Disease

TARGET	Population-Average Point Estimate	95% Confidence Interval	95% Prediction Interval	Number of Studies
Proportion of Recurrences Amenable to R0 Salvage in the Intensive Arm	0.35	(0.23, 0.49)	(0.10, 0.71)	8
Odds Ratio for R0 Salvage	2.1	(1.3, 3.3)	(0.8, 5.7)	8
Proportion of Recurrences Detected while	0.71	(0.62, 0.80)	(0.55, 0.84)	6

⁷ Strictly speaking, I first calibrated the linear coefficients to the *time-to-recurrent-disease* targets and the rate (odds ratio) of asymptomatic disease detection. I then calibrated the required parameters governing the resectable/not-resectable disease process (described below) using these initial probability distributions for the linear coefficients. I then re-weighted pairs of linear coefficients based upon their fit to all 3 targets.

Asymptomatic in Intensive Arm				
Odds Ratio for Asymptomatic Detection	3.5	(1.3, 9.2)	(0.32, 37.8)	6

The studies used are shown in the forest plots of Figures B.1-B.4 of Appendix B.

CALIBRATION RESULTS

The results of the calibration of the natural history model of recurrence are portrayed in two formats in Appendix B: a set of figures (Figures B.5-B.17) and Table B.2. For each location (rectal vs colon), stage (I-III), adjuvant therapy (FOLFOX, 5-FU + LV, or none), and period (old vs new) combination that was available, a figure depicts the *time-to-recurrent-disease* target (black) and the corresponding model-produced output (blue) using the parameter set with the highest probability mass (thus referred to as ‘Best Model’). Model-output based on the other included parameter sets is depicted in dashed brown lines. I remind the reader that each parameter set was evaluated by both the fit between the corresponding model-output (smoothed hazard function or KM survival curve) and the target and between model-produced surveillance output and the 4 meta-analytic targets of Table 2. I also note again that in each case the model-produced output is simulated using the surveillance schedule reported in the target source (if there was one). In some cases, the target resulted from a pooled analysis using patient level data, and thus this involved a mix of different follow-up regimens. In other cases, the regimen was only vaguely described in the protocol(s).

It should be noted that the y-axis scale varies across some figures. This might give the impression that, in some cases, the model-produced output does not fit the targets well. However, this is mostly an illusion of scale. In general, the model fit the targets well. For example, for stage II and III colon cancer patients treated with FOLFOX adjuvant therapy

(the fit of which might appear suspect), the model- and target-implied survival curves that correspond to the depicted hazard functions are within 1 and 2 percentage points, respectively, over the entire domain of the function when using the highest-probability parameter set. So for example, if the target-implied survival curve was 0.75 at some time point, the model-implied survival curve (using the best parameter set) would be guaranteed to fall within 0.74-0.76 or 0.73-0.77 for stage II and III patients, respectively.

Still, in general there was better fit between model-produced output and targets when the targets represented the disease experience of patients without intensive follow-up. This is likely due to three factors: uncertainty (due to non-reporting) regarding the follow-up adherence levels of included patients, a more widely dispersed scheduling of follow-up tests among actual trial patients than was assumed in the model, and differences in the degree of smoothing in hazard function estimation. Unfortunately, none of the *time-to-recurrent-disease* targets were accompanied by confidence bands in the literature, so it is difficult to judge to what degree there is any discrepancy between the curves that falls outside of the bounds of random variation. Though given the above-noted small deviation in survival curves, it is likely that such a discrepancy would have only a trivial impact on the results of analyses.

Table B.2 provides information about the distribution of sojourn times (in months) for each disease location, stage, adjuvant therapy, and period combination. As a reminder, the sojourn time is the length of time the disease is potentially detectable by surveillance tests yet pre-clinical. This is the period during which follow-up could detect asymptomatic disease and thereby increase the chances that salvage surgery could be performed. In particular, the table gives information about the mean, median, and first

and third quartiles of the distribution of sojourn times. Each pair of hazard functions (time-to-detectable- and time-to-clinical-disease) lead to a different distribution of sojourn times, and some pairs of hazard functions were better (higher-probability) than others. Thus, I calculated a weighted average for each moment/quartile of the distribution of sojourn times using the discrete probabilities derived from calibration as weights. These overall quantities (based on the weighted-average) are the primary result of the table and represent the model's best prediction concerning the mean/quartiles of the distribution of sojourn times. However, I also include (in parentheses) the range (over the different parameter sets) of each moment/quartile produced by the model. It should be noted that estimates are precise (with respect to Monte Carlo error) to the second decimal place, and thus Monte Carlo error can be ignored.

SURVEILLANCE TESTING AND DETECTION OF DISEASE

TEST SENSITIVITY

The key phenomenon the model seeks to accurately yet parsimoniously capture here is the ability (or lack-thereof) of extracolonic-focused surveillance tests (CT scans and CEA assays) to detect pre-clinical disease that is no longer dormant and has begun to develop. In the case of CT imaging, there is clear evidence of variation in the sensitivity of the test based upon the size of the largest lesion.^{84 85} Thus while it would be convenient to use an average sensitivity value in the model for a CT scan, this could potentially lead to misleading results. Moreover, such an average sensitivity value will inevitably be biased upwards with respect to its accuracy for small lesions since they will often not be detected and thus not included in studies of diagnostic accuracy. It was therefore necessary to model the size and thus growth of the largest lesion during the sojourn

period. While many patients will present with multiple lesions and possibly at multiple sites, the largest lesion would generally be the determinant of the sensitivity of a CT scan to detect asymptomatic disease. I thus model the disease as 1 lesion. However, the tendency for disease to colonize an organ and spread to other organs over time is relevant to the potential for salvage surgery. The resectability of disease is described in the next section.

About 65-75% of all recurrences will involve disease in the liver and or lungs^{4 45 86} so the model tried to capture the development of a hepatic or pulmonary lesion. Studies of CRC metastases in animals suggest that the tumor-development process is initially slow at small tumor volumes but then transitions to exponential-like growth (due to improved access to the circulatory system and the consequent resources) until it again slows down due to external limitations (e.g., runs out of space, insufficient access to resources, limited vascularization given the size of the tumor).⁸⁷ These observations suggest that a sigmoid-shaped parametric form is a reasonable choice to represent growth of the tumor.⁸⁸ In particular, I model tumor volume with a Gompertz growth curve (Equation 1)

$$V(t) = V_0 \exp\left(\frac{\alpha}{\beta} (1 - e^{-\beta t})\right) \quad (1)$$

For the purposes of modeling the (possible) detection of disease by CT imaging, the diameter of the tumor is more relevant than the volume. If we assume the lesion is spherical, there is a one-to-one relationship between the volume and diameter given by the equation: $V = \frac{4}{3}\pi(d/2)^3$.

The primary feature of the Gompertz growth model is that its relative growth rate $\left(\frac{\partial V(t)/\partial t}{V(t)}\right)$ decays exponentially. The rate of decay is controlled by the parameter β while

the parameter α represents the initial proliferation rate.⁸⁹ It is useful to define the limit of this function as t becomes large (the maximum achievable volume, or the carrying capacity), $V_\infty = V_0 \exp(\frac{\alpha}{\beta})$, where V_0 is the volume at time 0. Given information (assumptions) about V_0 and V_∞ , a classic Gompertz growth model depends only on (is fully specified by) the value of β .

In reality, the growth and progression of hepatic and pulmonary lesions during the sojourn period will be doubly heterogeneous: (a) the rate of growth and (b) the size at which disease becomes clinical will vary from person to person (and possibly within a person).⁸⁸ That is, even in the absence of surveillance testing, most lesions will not actually grow to their true carrying capacity since the disease will become clinically-indicated prior to that. However, there is insufficient empirical evidence to parse out the heterogeneity of these two processes. Thus, in order to make this modeling task tractable, I assume that all lesions become symptomatic at the same size, with a diameter of 5.5 cm. Moreover, I assume that the sojourn period begins (i.e., the disease becomes potentially detectable) when the largest lesion is 1 mm. Before that, the disease is assumed undetectable and so the size is irrelevant.

I selected 1 mm as the beginning of the sojourn period because a diameter of 1 mm was the smallest reported lesion I could find in the literature.⁹⁰ While the choice of a diameter of 5.5 cm as the size at which all lesions become symptomatic is necessarily arbitrary, it is not unreasonable. The vast majority (75%) of hepatic resections involve lesions of a smaller size⁹¹, and this is almost always true for pulmonary metastasectomies.⁹⁰ Moreover, while a lesion of this size might not always be large

enough to affect hepatic or pulmonary function to a degree that causes symptoms, it is likely the disease will have spread beyond the original organ by this point. Apparently, patients with clinically-indicated metastatic disease more often than not present with symptoms resulting from wide-dissemination of the disease rather than loss of hepatic or pulmonary function.⁹² Unfortunately, I was unable to find helpful information in the literature concerning the size of lesions in local-regional e.g., pelvic, recurrences, and so I am unable to evaluate the reasonableness of the selection of 5.5 cm for such patients.

To recap, the natural history model of tumor growth tracks the development of a metachronous lesion during its sojourn period as it grows from 1 mm in diameter to 55 mm in diameter and thus becomes symptomatic. The total time this process takes is determined by the time-to-clinical-disease simulation for each individual. As a consequence, at any given time in the simulation, different individuals with recurrence will have different size lesions. However, all lesions will be the same size after a given proportion of the patient's sojourn time. That is, all individuals with a recurrence will have the same size lesion after say 50% of their sojourn time has elapsed, however long that process takes. Thus in the simulation model, the size of a lesion at the time of a CT imaging study, and thus the sensitivity of the test, will be determined by what proportion of the patient's sojourn time has elapsed. The model was therefore in need of two empirical inputs: a (1) Gompertz curve characterizing the growth of the tumor from an initial volume with a diameter of 1 mm (0% of the sojourn time) to a final volume with a diameter of 5.5 cm (100% of the sojourn time) and (2) the sensitivity of CT imaging for different sized lesions.

For the first of these, I turned to previous empirical work⁸⁸ that used CT scans to measure the volume of two groups of hepatic metastases (a set of occult tumors and a set of larger, surgically-identified tumors) in patients at two or more time points. Using a Gompertz growth model, the authors estimated the average age of both groups of tumors for which a mean volume was available. I used these two mean tumor-volumes (implied diameters of 2.6 and 3.6 cm) and corresponding estimated average tumor-ages (2.3 and 3.7 years, respectively) to calibrate a Gompertz growth curve. To do this, I followed the previous authors and further assumed the tumor started as 1 cell with a volume of $1 \times 10^{-9} \text{ cm}^3$ and had a carrying capacity of 1000 cm^3 (a diameter of about 12.4 cm). These assumptions combined with the two data points implied a unique exponential decay parameter β and thus Gompertz growth curve. However, because I was only interested in the size of the tumor once it was potentially detectable, i.e., once it had at least a 1 mm diameter, I rescaled the time domain t to $\frac{t-\tau_1}{\tau_{55}-\tau_1}$, where τ_1 and τ_{55} are the times at which the calibrated Gompertz function was equal to an implied diameter of 1 mm and 55 mm, respectively. The resulting function (Figure C.1) is defined over the interval $[0, 1]$ and identifies a tumor volume (diameter) at any fraction of elapsed sojourn time.

The above function maps from the proportion of elapsed sojourn time to the diameter of the largest lesion. The next step was to construct a function that maps from the lesion diameter to CT sensitivity. I did this by fitting a LOWESS curve⁹³ to a set of data points taken from the literature giving the estimated sensitivity of a CT scan for different sized lesions (Figure C.2). I found two sources which provided estimates of average CT

sensitivity for different size-categories of hepatic lesions: a meta-analysis⁸⁵ and a single institution study⁸⁴ which categorized sizes differently than the meta-analysis. Similar sources were not available for pulmonary or local lesions. These values and the size-ranges to which they applied are given in Table 3. I assumed that the sensitivity of a CT scan for a lesion at 1 mm (55 mm) was 0.10 (0.98) because this was the smallest (largest) sensitivity reported in the literature.⁸⁵

The final step for the CT scan was to construct a composition of the two above-mentioned mappings: (i) from fraction of sojourn time elapsed to lesion diameter and (ii) from lesion diameter to CT sensitivity. The resulting function (Figure 2) identifies a CT sensitivity value for any fraction of elapsed sojourn time between [0, 1].

It is noteworthy that this implicitly assumes that an abdominal & thoracic CT scan is performed in the case of colon cancer patients and an abdominopelvic & thoracic CT scan in the case of rectal cancer patients. In real life, an abdominal CT scan would not fully visualize the lungs and so would not be able to detect pulmonary metastases. Thus, the effective sensitivity to detect any recurrence would be lower than assumed here. As previously noted, in the interest of simplicity and convenience, the model does not distinguish among different recurrence locations. Thus as currently implemented, ideally the model should only be used to simulate and evaluate the full CT combinations previously mentioned. Although it might provide for a reasonably approximate representation of the combination of abdominal ultrasound and chest x-ray or of an abdominal CT scan (ultrasound) performed at 6 or 12 months after primary resection. This is because over 75% of pulmonary recurrences present after one year.⁴

Table 3: Estimates of CT Sensitivity by Hepatic Lesion Size

AVERAGE SENSITIVITY (95% CI)	LESION SIZE RANGE (DIAMETER)	ASSIGNED SIZE (range for sensitivity analysis)	SOURCE
0.10 (0.05, 0.20)*	-	1 mm	Assumption
0.352 (0.218, 0.488)	1-10 mm	7.5 mm (6-9)	Ichikawa (2010) ⁸⁴
0.702 (0.605, 0.798)	11-20 mm	15.5 mm (14-17)	Ichikawa (2010)
0.888 (0.805, 0.970)	21-30 mm	25.5 mm (24-27)	Ichikawa (2010)
0.962 (0.910, 1.000)	31+ mm	40.5 mm (35.5-45.5)	Ichikawa (2010)
0.473 (0.401, 0.545)	1-9 mm	7 mm (5.5-8.5)	Niegel (2010) ⁸⁵
0.867 (0.776, 0.925)	10+ mm	36 mm (31-41)	Niegel (2010)
0.744 (0.687, 0.793)	All lesions	31 mm (26-36)	Niegel (2010)
0.98 (0.90, 1.00)*	-	55 mm	Assumption

* = Range for Sensitivity Analysis rather than 95% CI; CI = confidence interval

For the CEA assay, I constructed a similar function that mapped from the fraction of sojourn time that had elapsed to the sensitivity of a CEA test. Unfortunately, there was no information in the literature on the lesion-size-specific sensitivity of CEA tests. I therefore constructed the curve as follows. I assumed that the sensitivity of a CEA test was 50% greater than that of a CT scan at the beginning of the sojourn period as there is evidence of elevated CEA-levels anticipating micro-metastases that were missed by CT scans.⁹⁴ I further assumed that, by the end of the sojourn period (when the disease was clinical), 30% of recurrences would never have triggered elevated CEA levels (an assumption corroborated by the literature^{22 23}). Finally, I used an adjusted meta-analytic estimate of the sensitivity of CEA assays for colorectal cancer recurrence – 0.54 (95% CI: 0.46-0.62) - and assumed this applied to the same mean-sized lesion (31 mm

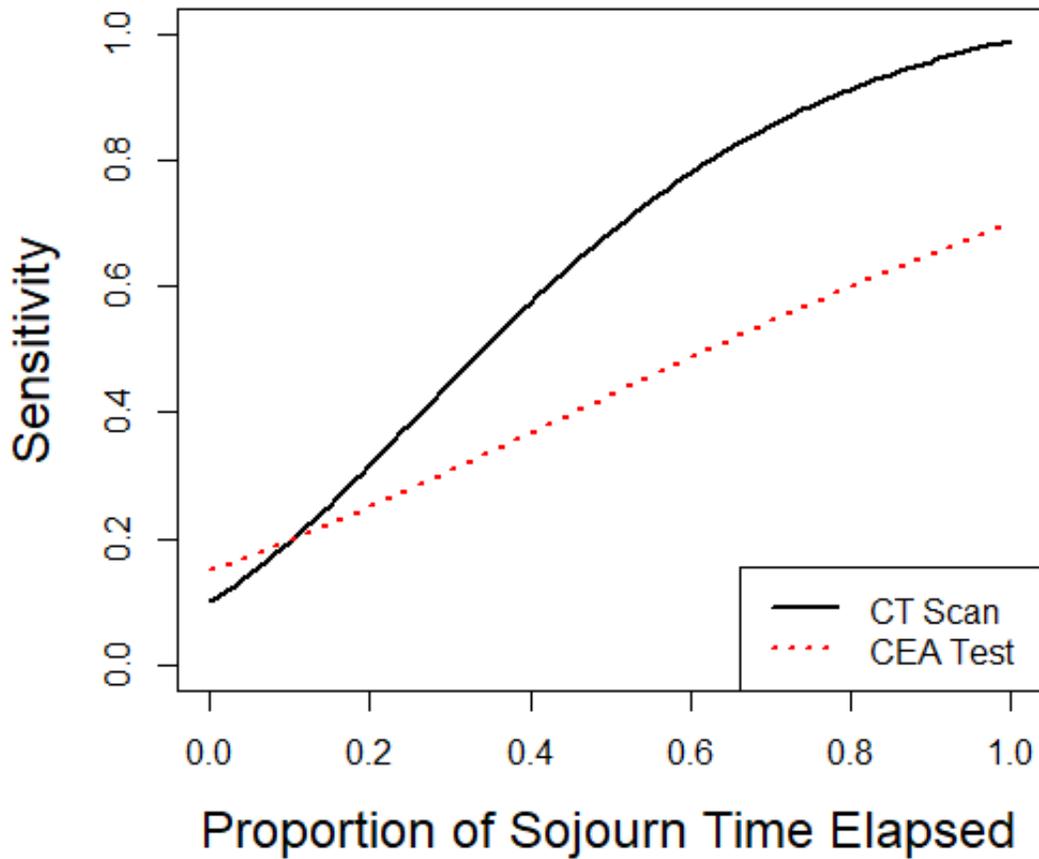
diameter) as in Niekel (2010). A Gompertz curve was fit using these three points and is depicted in Figure 2.

A recent systematic-review and meta-analysis²⁷ highlighted the high risk of bias in most reported estimates of the sensitivity of CEA due to the retrospective methods used to estimate this quantity and the questionable relationship between common study criteria for retroactively classifying CEA readings as detecting or missing disease and actual prospective clinical practice of interpreting CEA readings. To get the estimate given above, I thus combined the only study in that review that was assigned a low risk of bias⁹⁵ with a more recent high-quality secondary analysis²³ of the FACS trial which estimated the sensitivity of a single CEA assay test (with a repeat measurement to confirm elevated scores) used prospectively to determine appropriate follow-up. In both cases, estimates of the sensitivity were based on using the threshold of 5 µg/L for elevated CEA levels.

IMPLEMENTATION OF SURVEILLANCE & FALSE POSITIVES

The model simulates follow-up testing in all patients who are alive and ostensibly disease free. While guidelines suggest CEA tests and clinical visits (CT exams) every 3-6 (12) months, it is unlikely that patients would stick to that exact schedule even if they were 100% compliant. The model thus randomly assigns a follow-up time (time at which the test is performed) for each test using a truncated normal distribution with a mean equal to guideline time, a standard deviation of 3 months, and truncation points of +/- 1.5 months for CEA exams performed in the first 3 years and +/- 3 months for CEA exams performed after 3 years and for all CT exams.

Figure 2: Sensitivity of a CT Scan and CEA Test vs. Elapsed Sojourn Time



A schematic representation of the implementation of surveillance in the microsimulation model is given in Figure 3. Any surveillance tests are necessarily wasted in patients who are disease free (cured) and in patients who have micro-metastases that, while currently undetectable, will eventually begin to grow and spread. Both of these types of patients can fall victim to false-positive findings however. Other than the possibility of false positives, the model assumes there are no potential personal harms associated with noninvasive follow-up testing as there would be with say endoscopic follow-up tests. While the exposure to ionizing radiation from an abdominopelvic and

chest CT scan is non-trivial, the lifetime risk of developing cancer from radiological imaging is very small among older patients (like CRC patients) who generally don't live long enough for subsequent disease to develop.⁹⁶

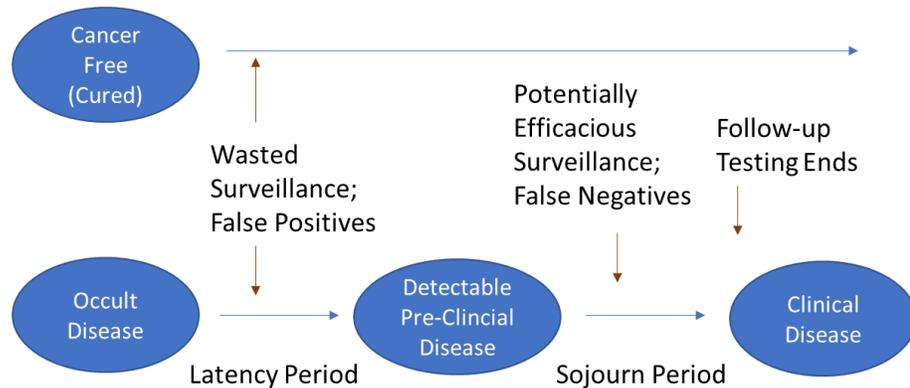


Figure 3: Schematic Representation of the Model Implementation of Surveillance

For patients who have detectable but pre-clinical disease, a given surveillance test has a chance of detecting their disease equal to the sensitivity associated with the respective modality (CT scan or CEA test) and the proportion of their sojourn time that has elapsed (see Figure 2). Pseudo-random simulation of a Bernoulli variable is used to determine whether a patient's recurrence is actually detected or not by the given test. Once the disease is detected, be it clinically-indicated or surveillance-detected, follow-up testing is discontinued and the patient is treated for recurrent disease (next section).

In the case of a CT scan, a true positive result means the patient has diagnosed recurrence and proceeds to treatment. It is unclear from the literature how commonly laparoscopy (or thoracoscopy for pulmonary findings) or imaging-guided biopsies are needed, but it is likely that some proportion of patients require these before surgical resection is considered due to equivocal imaging results. I therefore arbitrarily assumed 5% of patients with a true-positive CT scan get a laparoscopy (or thoracoscopy) and 5%

get a CT-guided fine-needle aspiration (FNA). This is simply a cost added to follow-up testing and does not influence a patient's trajectory through the model. Moreover, the model currently assumes a pelvic MRI is performed after one of every four true-positive CT scans identifying recurrent disease in rectal cancer patients. The point of this is to rule out false-positive findings in the pelvis and better determine the appropriateness of salvage surgery.^{97 98} As with the laparoscopy/FNA, there is no simulation associated with the MRI; it simply represents a cost. The one-in-four value was selected because approximately 20-25% of recurrences seen in rectal cancer appear in the pelvis.^{4 45} In its current form, the model does not implement a PET scan or PET/CT combination scan after a true-positive CT scan as might be used to help identify widely-disseminated (extrahepatic abdominal) disease to avoid unhelpful surgery.^{99 100} This is because it is unclear how common this procedure is in actual clinical practice. Moreover, the model currently allows for failed surgical resection.

In the case of a CEA assay, a true positive represents two consecutive elevated readings ($> 5 \mu\text{g/L}$) 2 weeks apart and leads to a chest and abdominal (abdominopelvic) CT scan. The sensitivity of a CEA test and subsequent CT test are assumed independent conditional on the fraction of elapsed sojourn time. The CT scan identifies the disease with probability equal to the previously described sensitivity, and the patient proceeds to treatment (with the added cost of a pelvic MRI for 25% of rectal cancer patients and a laparoscopy or CT-guided FNA for 10% of all patients). If, however, the CT scan fails to identify the disease, no further tests are performed until there is another positive routine CT or CEA test. That is, the patient's disease is missed (a false negative). In its current form, the model does not implement a PET Scan in these circumstances (elevated CEA

levels but negative CT imaging) as might be considered^{94 98} again because it is unclear that this is representative of general clinical practice in the US.

A major cost of follow-up testing is the apparently relatively common occurrence of false-positives.^{23 101 102} The costs of false-positives are wasted resources and potential psychological harms to the patient. Currently, the model does not include the latter as reliable information is not available in the literature. For false-positives, there were two cases to consider: (i) non-malignant findings on a routine CT scan and (ii) elevated CEA-levels and a subsequent CT scan. For a routine CT scan, a patient has a false-positive with probability equal to 1-specificity. For specificity, I use an estimate from the hepatic imaging literature – 0.949 (95% CI⁸ 0.929-0.963)⁸⁵ - because better information exists, but comparable results have been reported for multidetector pelvic CT scans¹⁰³. I assume that a false-positive CT scan requires further follow-up. In particular, I arbitrarily assume that patients get one of four options in equal proportion: (1) repeat CT scan (of the abdomen or chest), (2) an MRI (of the pelvis or abdomen), (3) a CT-guided FNA, or (4) a laparoscopy or thoracoscopy.

In the case of CEA, I use a specificity of 0.93% (95% CI: 90.6–95.3%), which was estimated using a threshold of 5 µg/L.^{23 95} A false positive leads to a CT scan (just as a true positive does). However, the model uses a different specificity for a CT scan undertaken to discover suspected recurrence (because of elevated CEA levels) – 0.74 (95% CI: 0.67-0.82) - than for a routine CT scan as there is evidence of a higher

⁸ The 95% CIs are used for a sensitivity analysis

incidence of false-positives in this context.¹⁰⁴ False positive CT findings lead to additional investigations as described above.

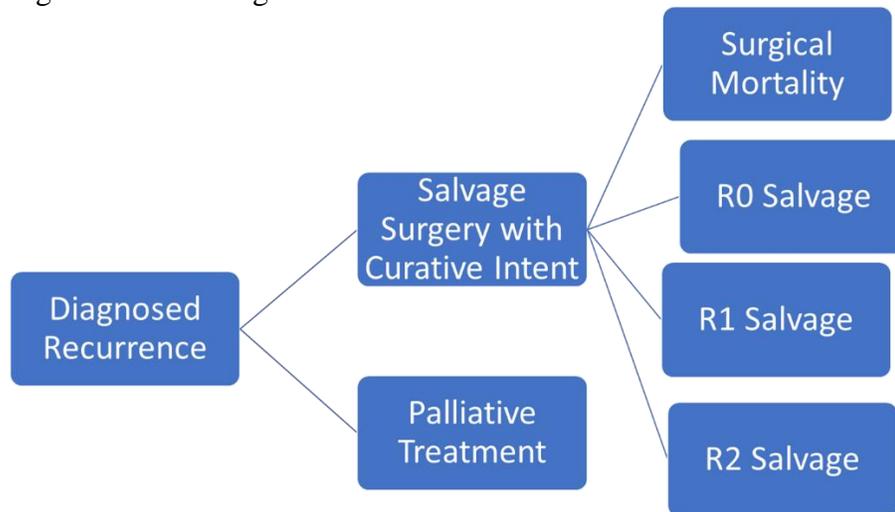
TREATMENT OF RECURRENCE

Patients with recurrence are either treated with curative intent or with the intent to prolong life and reduce symptoms. For convenience, I refer to the latter as palliative treatment. Figure 4 depicts a flow diagram for patients diagnosed with recurrence. The primary modeling task is the determination of which form of treatment a patient gets. In order for a recurrence to be a candidate for salvage surgery with a curative intent, in general the disease must be isolated (in one organ) and, in the case of liver and lung metastases, sufficiently isolated and concentrated within the organ to allow the organ to function after resection. As time progresses, the disease is likely to further colonize the initial organ (e.g., lesions increase in volume, more lesions develop, disease spreads to another lobe of the liver) and spread to other parts of the body. Thus, over the course of the sojourn period, the chances the disease is salvageable will decrease dramatically.

Mimicking the above-described model of the growth of the largest lesion, I modeled the trajectory of the chances the disease is unresectable upon detection (during the sojourn period) using a Gompertz function. I assumed that around 25%⁵ and 35%¹¹ of clinically-indicated, isolated local-regional recurrences in colon and rectal cancer patients are salvageable, respectively, and that isolated local-regional recurrence constitutes about 15%⁵ and 20%^{4 11} of all recurrences among colon and rectal cancer patients, respectively. I further assumed that only 5% of distant metastases are salvageable once symptomatic. Together these assumptions entail a salvage rate of roughly 8% and 11% for clinically-indicated recurrences in colon and rectal cancer patients, respectively, i.e., at the end of

the sojourn period. It is unknown what proportion of recurrences are salvageable at some point during their sojourn period, but it is likely that some never are. I therefore assumed that 10% of recurrences are never resectable. Thus the probability a recurrent disease is unresectable if discovered at the beginning of the sojourn time is 10%.

Figure 4: Flow Diagram of Treatment of Recurrence

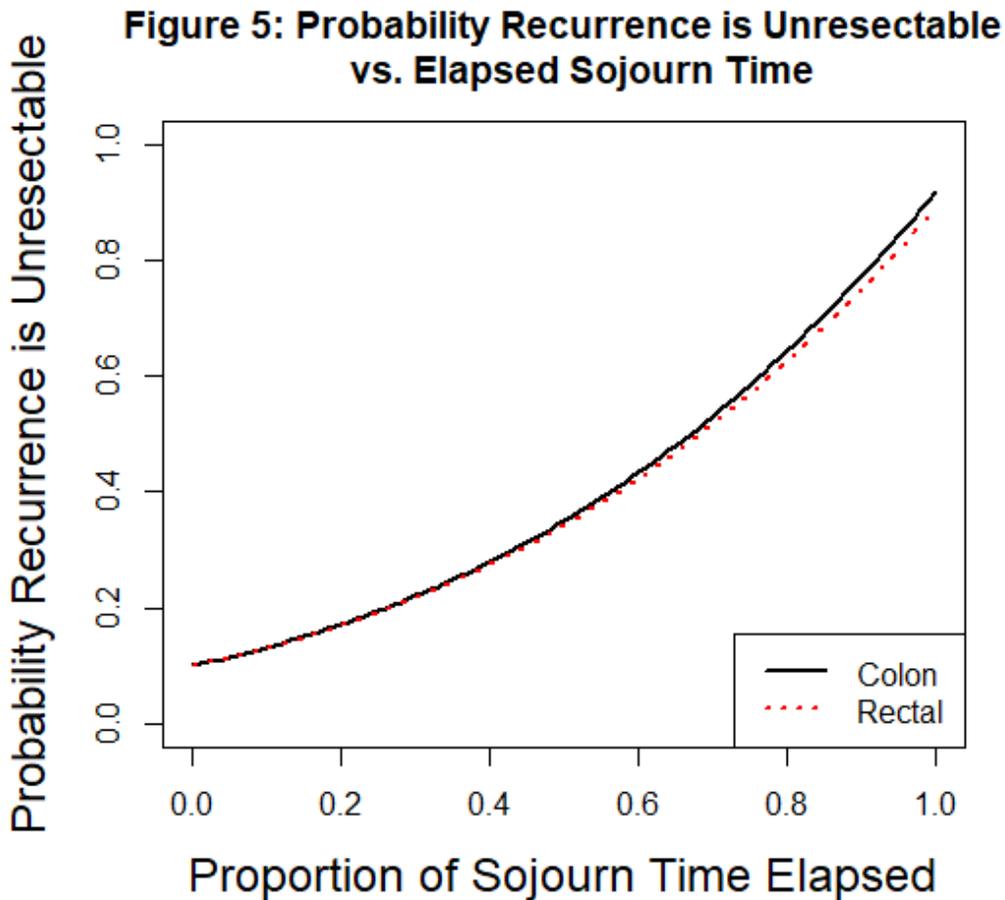


In order to estimate a best-fit Gompertz curve, I calibrated a third point at an arbitrary timepoint of 70% of the sojourn period. This proceeded as follows. Using a mix (equal proportion) of colon and rectal cancer patients, stage II and III patients, and, for colon cancer patients, no adjuvant chemotherapy and adjuvant chemotherapy (5-FU/LV and FOLFOX), I selected a reasonable starting value for the unknown parameter and simulated two cohorts of patients for five years: one undergoing the ‘intensive’ and one undergoing the ‘minimal’ follow-up regimen used in the calibration of the natural history model mentioned above (Table B.1). For each subgroup of patients, e.g., stage III colon treated with FOLFOX, I used the highest probability parameter set, i.e., set of time-to-detectable-disease and time-to-clinical-disease hazard functions. As mentioned in footnote 4, the initial probability distribution for the different pairs of parameter sets

(hazard functions) was based off an initial calibration exercise using only the *time-to-diagnosed-recurrence* targets and the two meta-analytic targets (Table 2) associated with detection of asymptomatic recurrence (and thus not those associated with salvage rates). For 44% (56%) of patients undergoing a salvage surgery, I used the new (old) distribution of surgical outcomes (Table 4: explained below). This was based on the fact that 44% (56%) of the meta-analytic weight for the salvage targets came from trials conducted after (before) 2000. The main difference between the new and old distributions of surgical outcomes is that the proportion of salvage surgeries which end in R0 resection has increased over time, increasing the expected benefit of attempted salvage surgery. For each simulation, I evaluated the model-produced intensive-arm proportion of recurrences undergoing R0 salvage and the odds ratio comparing the two arms. In particular, these two model produced statistics were evaluated with the corresponding meta-analytic predictive-distribution likelihoods described above. I again used R's simulated annealing (available in the function `optim`) to find the parameter (the proportion of recurrences that are unresectable after 70% of the sojourn time has elapsed) that maximized the product of the two likelihood scores.

The resulting function relating the probability that a diagnosed recurrence is unresectable with the proportion of the sojourn time that has elapsed is depicted in Figure 5. A patient with diagnosed recurrence receives palliative treatment with probability equal to the chance the disease is unresectable and receives curative resection with probability equal to one minus the probability the disease is unresectable. For patients treated palliatively, the model simply simulates a time-to-cancer-death variate as described in the next section. Costs for such a patient are discussed in Chapter 4.

For patients treated surgically with a curative intent, there is a small risk of 30-day surgical-related mortality (2%) and a larger risk of surgical-related morbidity. The 2% figure was selected as a reasonable balance between the higher rate reported for hepatic resections (2.5%)⁶⁵ and the lower rate reported for pulmonary resections (0.9%)⁶⁶ in a population setting. This value applies to patients under 75. For patients 75 and old, there is evidence of a higher 30-day mortality rate (5%).¹⁰⁵ Surgical-related morbidity is discussed in Chapter 4.



Patients who survive salvage surgery are then assigned a surgical outcome. These include a terminated resection due to the discovery of widely-disseminated disease or an

incomplete resection with residual macroscopic disease that is visible to the surgeon (which I group as an R2 resection), microscopically-positive surgical margins (R1 resection), and macroscopically- and microscopically-clear margins (R0 resection). The distribution of these three surgical outcomes differs for metastasectomies (generally the liver and lungs) and surgical resection of local-regional recurrence (isolated pelvic disease for rectal cancer or isolated abdominal recurrence in colon cancer). The former is more favorable. This is likely due to the constricted space of the pelvis and the potential for disease to be appended to nerves and/or bones.⁶¹ Since a sizeable proportion of local-regional recurrences that present as clinical disease are salvageable^{5 11 102}, and since symptomatic metastatic disease is essentially never salvageable⁶, the model uses the distribution of surgical outcomes associated with metastasectomies for surveillance-indicated disease and the distribution associated with local-regional disease for clinically-indicated disease. These are given in Table 4. They are taken from an (author-updated) meta-analysis. It is of note that both distributions have improved over time.

Table 4: Distribution of Surgical Outcomes among Patients Undergoing Salvage Surgery

DETECTION METHOD	Probability of R0 Resection	Probability of R1 Resection	Probability of R2 Resection	Sources
(New) Clinically-Indicated Disease	0.605	0.282	0.113	61 106
(Old) Clinically-Indicated Disease	0.436	0.204	0.360	
(New) Surveillance-Indicated Disease	0.822	0.149	0.029	91 107-118
(Old) Surveillance-	0.596	0.108	0.296	

Indicated Disease				
----------------------	--	--	--	--

New = Study published after 2000; Old = Study published before 2000.

DEATH FROM CANCER

The model simulates the life-course of patients from immediately after curative surgery for the primary tumor to death, be it from cancer or other causes. Patients who do not experience a recurrence (those who are cured) die from other causes. The rate at which this happens depends upon the application of the model. For the purposes of Chapter 4, I simulate death times from the most recent (age- and sex-specific) US lifetables available from the National Center for Health Statistics¹¹⁹ in the Center for Disease Control and Prevention (CDC). Patients who suffer a recurrence can die from cancer or other causes, whichever happens first. The rate at which patients die from cancer depends upon the treatment they receive: palliative treatment or resection with a curative intent. Patients treated palliatively or who receive an R2 attempted salvage surgery are considered to have terminal disease and face a grim outlook. Patients who receive an R1 or R0 salvage surgery can be cured, but most will eventually suffer a recurrence and die from cancer. The re-recurrence process is not explicitly modeled. Instead patients who are uncured simply die a cancer-specific death at a time simulated from an appropriate distribution described below. In what follows, I describe the time-to-cancer-death processes for patients treated curatively and for patients treated palliatively.

PALLIATIVELY TREATED PATIENTS

Patients treated palliatively face an extremely high mortality rate and so will almost all die from cancer before they would have otherwise died. The length of their remaining lifetime (in a world where the only mortality risk is CRC-specific death) is simulated

from a calibrated survival curve adjusted for individual-specific frailty terms. I again used a discretization approach to simulate from the survival curve as it was significantly faster than rejection sampling and lead to only trivial bias. Calibration of the time-to-death frailty variance and the correlation with other heterogeneity terms is discussed below. The baseline survival curve was taken from a meta-analysis of randomized control trials of systemic chemotherapy for terminally-ill CRC patients. The curve characterizes the survival experience of a mix of patients who were deemed to have unresectable metastatic disease, some with clinical disease and some with surveillance-detected disease. The latter patients will tend to live longer from the time of diagnosis simply by virtue of the lead time. For the purposes of the simulation model, the baseline survival curve thus needed to be adjusted to ensure patients with surveillance-detected disease were not penalized for earlier detection of disease.

This time-to-death (from terminal disease) process is implemented in the microsimulation model as follows. Patients with clinically-indicated, unresectable disease are given a remaining (CRC-specific) lifetime simulated from the adjusted (calibrated) survival curve. Patients with surveillance-detected disease are also given a remaining (CRC-specific) lifetime simulated from the same survival curve but which is added onto their lead time – the time interval between detection of their disease and the onset of symptoms. The latter is available since the microsimulation model uses a discrete-event framework. Thus all such patients at least live to the time at which their disease would have become clinical had it not been detected asymptotically. In the current form of the model, there is no survival advantage conferred from asymptomatic detection of unsalvageable disease.

Two different survival curves were taken from the literature and calibrated. The first was from a meta-analysis using patient-level data and represented the disease-experience of terminally-ill patients treated with single-agent fluoropyrimidines (5-FU or capecitabine).¹²⁰ The second was a weighted average of a meta-analytic curve using patient-level data¹²⁰ and a new randomized control trial¹²¹ and represented the disease-experience of terminally-ill patients treated with the double-agent regimen of FOLFOX . The latter is more common in contemporary clinical practice. In each case, overall-survival curves were digitized using PlotDigitizer and implemented in R using spline interpolation.

Cancer-specific survival (CSS) curves were backed out from these overall-survival curves using the following approach. Details of the theory behind this approach can be found in Chapter 2 of Klein and Moeschberger.¹²² I assumed that death from cancer and death from other causes were independent competing risks. Under the assumption of independent competing risks, marginal and cause-specific hazard rates are the same. Using lifetables from the Human Lifetable Database, a marginal other-cause mortality hazard rate (equal to the cause-specific hazard rate by the assumption of independence) was estimated for the period covered by the overall-survival curve. The definite integral of the product of this hazard rate and the overall-survival function provides an estimate of the crude cumulative incidence of other-cause mortality. Moreover, with independent competing risks, the negative integral of the ratio of the derivative of this crude cumulative incidence (for other-cause mortality) to the overall-survival function provides an estimate of the cumulative hazard function for cancer-specific mortality in the absence of competing risks. The target survival curves were only available for up to 5-7 years.

Because the cumulative hazard functions were roughly linear at the end of that period (signifying a constant hazard rate), I used linear extrapolation to extend the function to 10 years. Finally, this was easily converted to the cancer-specific (net) survival function (in the absence of competing risks).

The above-described CSS functions served as data targets for the next step which required calibration. The CSS functions were adjusted with a calibrated HR (assuming proportional hazards) to better represent the mortality experience of a cohort of patients with clinically-indicated disease. Thus each patient's time-to-death from unresectable disease is the sum of a simulation from this distribution and their lead time (0 for clinically-indicated patients). The value of this HR was selected so that, among a cohort of patients undergoing intensive surveillance, the mortality experience of the subgroup of patients with unresectable terminal illness (both clinically- and surveillance-indicated) best mimicked the target curves. The HR was calibrated in the presence of heterogeneity (adjustment of baseline risk by personal frailty terms).

Figure 6 depicts the two target CSS curves and the two calibrated CSS curves for patients with clinically-indicated unresectable disease. For symptomatic palliative patients treated with single-agent fluoropyrimidines (referred to as 5-FU in Figure 6), the median cancer-specific survival is estimated to be 10.4 months. This is consistent with the median overall survival of 10.8 months observed in the FACS trial among patients treated palliatively in the minimal surveillance arm. Almost all of these patients presented with symptomatic disease. It is also consistent with the median relative survival of 9.2 months observed among recent stage IV patients in the Surveillance, Epidemiology, and End Results (SEER) registry who underwent no surgery for primary disease (for reasons

other than comorbidity). Such patients are likely to have presented with clinical, unresectable disease as otherwise the primary tumor should have been resected. Note that this statistic excludes patients who did not undergo resection due to other comorbidities because such patients would not be included in surveillance cohorts. For symptomatic patients treated palliatively with FOLFOX, the median survival was 15.8 months.

PATIENTS TREATED WITH A CURATIVE INTENT

In the simulation model, patients treated with an attempted salvage surgery that ends in an R2 (incomplete or stopped) surgical resection receive no survival benefit from the procedure and are considered equivalent to palliatively-treated patients. Patients who undergo an R0 or R1 margin surgical resection are considered to have been treated with a curative intent. They have a chance of cure. Following other authors in the literature, I assumed that 10-year survival after salvage surgery indicates cure.^{123 124} Ten-year survival rates for patients with an R0 resection were taken from an author-updated meta-analysis.

There is evidence of a better prognosis among patients who present initially with a node-negative disease (stage I/II) than for patients with node-positive disease (stage III).^{91 125} I therefore backed out cure rates (Table 5) for each subgroup of patients using an estimated meta-analytic OR (0.46; 95% CI = 0.26-0.79) of the chances of surviving to 10 years comparing node-positive disease to node-negative disease¹²⁵ and assuming that about 60% of salvages involved patients with node-positive primary disease^{91 125 126}. There was also evidence of an increase in the cure rate over time, and so I calculated rates for publications prior to 2005 and for those after 2005. Table 5 also provides implied

95% credible intervals (used for sensitivity analyses) for each cure rate based upon the compounded uncertainty of potentially multiple parameters.

Figure 6: Cancer-Specific Survival for Unresectable Recurrent Disease

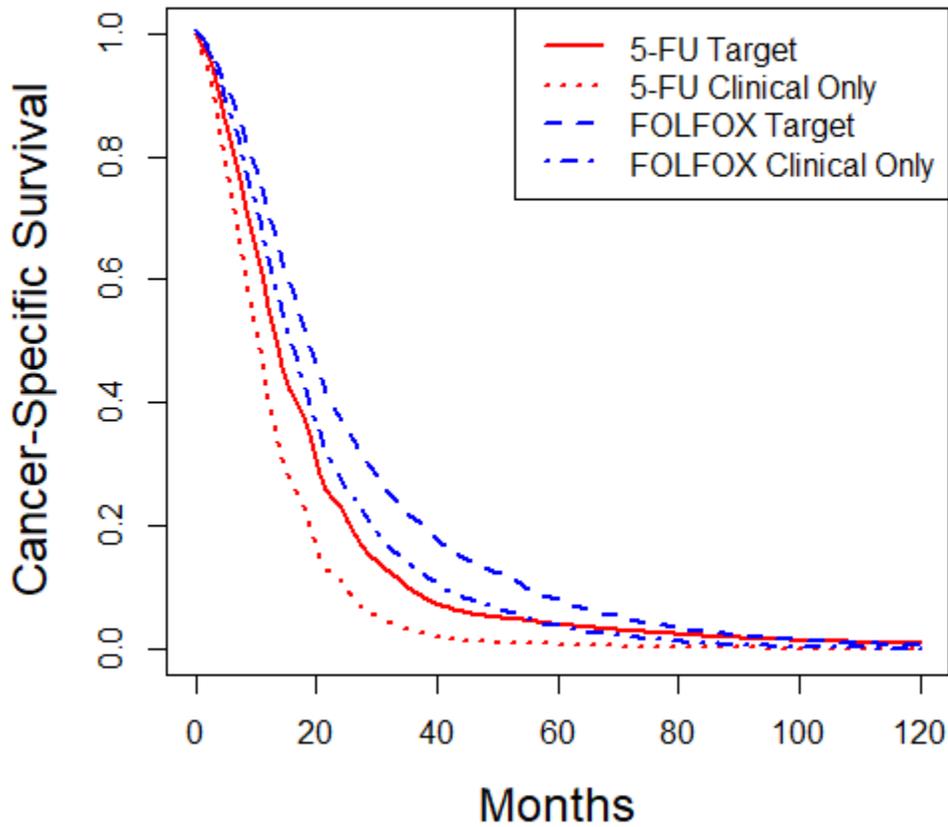


Table 5: Cure Rates after Salvage Surgery

PERIOD	PATIENTS	CURE PROBABILITY	IMPLIED 95% CREDIBLE INTERVAL	SOURCES
After 2005 (New)	R0 Salvage Stage I/II	0.427	(0.180, 0.717)	65 66 91 123 125
	R0 Salvage Stage III	0.255	(0.090, 0.544)	
Before 2005 (Old)	R0 Salvage Stage I/II	0.351	(0.219, 0.503)	
	R0 Salvage Stage III	0.199	(0.110, 0.325)	

After 2005 (New)	R1 Salvage Stage I/II	0.141	(0.000, 0.274)	123 124
	R1 Salvage Stage III	0.070	(0.000, 0.154)	
Before 2005 (Old)	R1 Salvage Stage I/II	0	0	
	R1 Salvage Stage III	0	0	

Although the evidence is somewhat equivocal, a shorter disease-free interval (time from resection of primary tumor to recurrence) appears to portend a worse prognosis after salvage surgery for both hepatic⁶⁵ and pulmonary²⁹ metastases. This is presumably due to a more aggressive biology of the disease, and this phenomenon is implemented in the model via correlated heterogeneity terms. However, the disease-free interval evidently does not predict 10-year survival rates.¹²⁵ Cure rates are therefore not adjusted for the disease-free interval.

Less evidence was available with regards to long-term survival among patients who receive an R1 resection. It was traditionally thought to rule out the possibility of cure.¹²⁴ While some recent sources have found encouraging results at 3 or 5 years when patients are also treated with (neo)adjuvant chemotherapy^{112 123}, most studies did not include sufficient follow-up to estimate 10-years survival rates. I therefore used the only estimate available¹²³ and assumed that in contemporary practice 10% of patients undergoing R1 resections will be cured.

Patients who undergo an R0 or R1 salvage surgery but are not cured will, in the absence of other causes of mortality, die from re-recurrent disease within 10 years. This appears to happen at a very similar rate among patients who undergo hepatic⁶⁵, pulmonary⁶⁶, and pelvic resection⁶⁷. To simulate this process among the uncured, I

constructed four CSS curves from the literature with which to simulate a time-to-cancer-death. The four curves were based on two factors: margin status (R0 vs. R1) and nodal status of the primary disease (stage I/II vs. stage III). Again, the effect of disease-free interval was implicitly modeled by adjusting the baseline risk of death (among the uncured) by a personal mortality-related frailty term that is strongly correlated with other time-to-disease processes. For a baseline curve, I used a digitized 5-year cancer-specific survival curve for patients who underwent salvage surgery for hepatic metastatic disease in a community setting in 2005 or after.⁶⁵ From this I constructed the four separate curves using (a) knowledge of the proportion of patients with stage III disease and who underwent R0 resection and (b) meta-analytic all-cause mortality hazard ratios. The latter were 2.02 (95% CI: 1.65-2.48)¹²⁶ - comparing R1 to R0 resections – and 1.6 (95% CI 1.4-1.7)^{91 126} – comparing stage III to stage I/II disease.⁹ The four curves were then extended to 10 years by linear extrapolation of the cumulative hazard so that the 10-year cancer-specific survival rates were equal to the respective cure rates. They are depicted in Figure 7. Finally, survival probabilities were renormalized to apply to only those who were uncured, i.e., survival probability of 0 at 10 years.

HETEROGENEITY VARIANCES AND CORRELATION

The correlation among log-frailty terms and the variances of each term (time-to-detectable-disease, time-to-clinical-disease, and time-to-cancer-specific-death) were calibrated¹⁰ to target HRs from the ACCENT database comparing the all-cause mortality

⁹ The actual implemented HRs were adjusted via calibration until model output (simulated in the presence of heterogeneity) exhibited a cancer-specific mortality hazard ratio equal to 2.02 and 1.6, respectively.

¹⁰ In the case of time-to-detectable disease and time-to-clinical disease terms, they were re-calibrated

hazard by year of recurrence.⁵⁶ The targets, model-output, and calibrated parameters are given in Table 6. Since the relationship between disease-free interval and survival after recurrence appears to be primarily a feature of the natural history of recurrence among patients with stage III colon cancer⁵⁶, I calibrated the heterogeneity parameters using stage III colon cancer patients treated with one of FOLFOX, 5-FU/LV or surgery alone. However, the model uses the same parameters for all forms of disease. Moreover, since nearly 80% of patients described by the target HRs were randomized to treatment prior to 1993, I calibrated the heterogeneity parameters using a cohort of patients undergoing intensive follow-up with CEA testing only.

Figure 7: Cancer-Specific Survival after Curative Resection by Margin Status and Stag

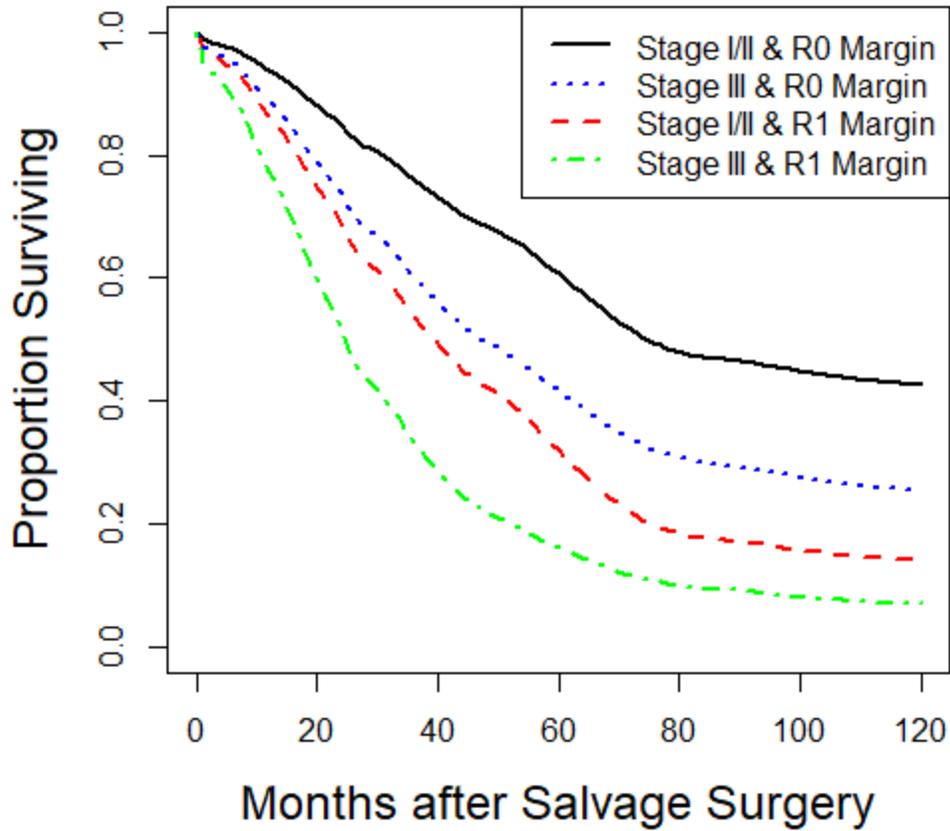


Table 6: Targets and Model Output for Heterogeneity Parameters

TIMING OF RECURRENCE	TARGET HR ⁵⁶ (95% CI)	MODEL-PRODUCED HR	CALIBRATED PARAMETER VALUES
1-12 Months	1 (reference)	1 (reference)	TTDD Var = 0.210 TTCD Var = 0.10 TTCSD Var = 0.11 Correlation = 0.71
13-24 Months	0.82 (0.77, 0.88)	0.79	
25-36 Months	0.69 (0.63, 0.75)	0.67	
37+ Months*	0.60 (0.54, 0.69)	0.60	

Table 6 gives all-cause mortality HRs associated with the disease-free interval. Patients who present with disease later (e.g., after the first year) face a reduced mortality hazard.

*I included recurrences up to 7 years after the primary resection.

CI = Confidence Interval; TTDD = time-to-detectable-disease; TTCD = time-to-clinical-disease; TTCSD = time-to-cancer-specific-death; Var = variance.

CHAPTER 3: AN EVALUATION OF THE LOGIC AND EVIDENCE BEHIND INTENSIVE FOLLOW-UP

INTRODUCTION

It is this author's contention that, while some skepticism towards the value of intensive surveillance is certainly justified, the conclusions of the above critics and calls for a radical reassessment of CRC surveillance practices are premature. As discussed above, critics of intensive follow-up have generally explained the negative results of recent trials by questioning the existence of a survival benefit from aggressive use of salvage surgery. However, there is reason to take seriously the theory that, when performed on patients with isolated disease, it is effective at improving survival and that this level of efficacy is attainable even in the context of salvage rates as high as 30-50%.

Many large case-series reports and systematic reviews have documented significant long-term survival (5/10 year overall survival of 25-65%/15-35%) after R0 salvage of hepatic^{91 123 125}, pulmonary²⁹, and local (pelvic)^{67 127} recurrences. In the case of hepatic^{65 128} and pulmonary⁶⁶ metastectomy, these results have been reproduced in a community-practice setting (as opposed to just at centers of excellence). Moreover, multiple studies have shown that outcomes have improved with time^{65 66 91 127} as case-volume has increased. While these case-series might be dismissed as involving highly-selected patients, these results have been reproduced in multiple modern cohorts of unselected patients undergoing intensive follow-up including routine CEA testing and imaging studies (Table 7). While this obviously does not definitely establish the benefit of salvage surgery, it shows that the impressive survival results seen in case-series cannot be

dismissed as due to selection of a small percentage of patients with non-aggressive disease.

The skeptical reader might question whether the results from the non-randomized, retrospective cohort studies identified in Table 7 (all but the FACS trial) are favorable because of selection of the patients involved in the analyses. However, there is little reason to suspect this. In general, the studies included all or almost all patients who underwent surgical resection for primary CRC (possibly just stage II-III patients and possibly just rectal) at the respective institution during the enrollment period. Arriola et al., (2006) and Kobayashi et al., (2007) included all patients of the included stages (II-III and I-III, respectively) treated at their center over the relevant period. The report by Ikoma et al., (2017) did the same with the exception of excluding patients who underwent emergency surgery for their primary tumor. Since the study focused on rectal cancer patients only, this might include around 10% of cases.¹²⁹ It should be noted that emergency surgeries are necessitated by obstruction and/or perforation and are more likely to lead to R2 or at least R1 surgical margins for the primary disease^{130 131}, and so it is likely that many of such patients would be excluded from a randomized trial evaluating intensive follow-up as they would not be considered potentially cured. Anyhow, this is likely less selective than both the FACS and GILDA trials which excluded patients who presented with (were found to have) recurrence prior to completing their adjuvant therapy.

Laubert et al., (2010) included all patients but divided the sample into three cohorts based upon the surveillance they actually received: ‘intensive’, ‘minimal’, and ‘none’. The minimal group contained patients who received less than 70% of the tests included in

the institution's protocol. The result presented in Table 7 is for the intensive arm patients only. It is unclear what degree of compliance was typical among the minimal arm participants, but it could plausibly be argued that the minimal arm participants should be included with the intensive arm participants since lack of adherence appears in some trials and is likely a reality of follow-up. In this case, the salvage rate of 30% would drop to 27%. However, it seems inappropriate to include the no-surveillance cohort given likely explanations of no follow-up are that patients were old or had serious comorbidities. It is unlikely that these patients would have been included in a RCT.

More importantly, the results for the FACS trial intensive arms are comparable with these results (Table 7). Two recent studies based on large community-practice case-series reported a median survival after hepatic⁶⁵ (52 months) and pulmonary⁶⁶ (51 months) salvage surgery (when limiting results to 2005 and after) that was nearly identical to that observed after curative resection in the FACS trial (52.3 months). The 8 patients treated in the minimal follow-up arm (1 CT only) had the most favorable survival experience (median OS = 76.9 months) likely due to the necessity that such patients had slow-moving disease since otherwise they probably would not have been candidates for salvage surgery. The 21 patients in the routine CT and CEA arm had the second best prognosis (median OS = 58.7 months) followed by the 19 patients in the routine CEA-only (+1 CT scan) arm (median OS = 51.2 months) and then the 28 patients in the routine CT-only arm (median OS = 43.6 months). The median OS was not given for the 3 intensive arms alone (excluding the 8 patients in the minimal arm), but the value was 52.0 months among the two routine CT arms (which includes the patients from the worst-performing arm: the CT-only arm).

This suggests the benefits of salvage surgery can be reproduced among a relatively high percentage of patients who suffer a recurrence in the context of intensive follow-up. Moreover, if selection of patients with less aggressive disease was responsible for this promising survival statistic observed in the FACS trial among patients undergoing curative resection, we would expect to observe a more favorable survival experience among palliatively-treated patients (non-curatively) in the FACS minimal follow-up arm than among the same group of patients in the intensive arms. This is because the patients with the less aggressive disease would have been selected out of this group in the intensive arms but not in the minimal group. That is, all the slowest-moving cases of disease would have been selected into the ‘surgically-resected’ group and the remaining cases would be the most aggressive. Thus we should see a worse prognosis in the intensive arm group of patients who did not undergo salvage surgery. However, the opposite was observed. In the 3 intensive arms, median survival after recurrence among those treated non-curatively ranged from 13-22 months while the same figure for the minimal arm was 10.6 months.

Table 7: Salvage Rates and Survival After Recurrence in Unselected Cohort Studies

STUDY	INDEX YEARS	COHORT N	RECURRENCE N (%)	SALVAGEN (%)	5 YEAR & MEDIAN OVERALL SURVIVAL AFTER RECURRENCE
Ikoma ¹³² (2017)	1993-2008	735	151 (20.5%)	70 (46.4%)	5 Yr. = 51% Median = 61.2 months
Laubert ¹³³ (2010) Overall	1990-2006	1469	211 (14.4%)	64 (30.0%)	5 Yr. Hepatic = 47% 5 Yr. Pulmonary = 66% 5 Yr. Local = 57%
Kobayashi ²¹ (2007)	1991-1996	5230	906 (17.3%)	379 (41.8%)	5 Yr. Hepatic = 45% 5 Yr. Pulmonary = 48% 5 Yr. Local = 30%
Arriola ¹³⁴ (2006)	1993-1999	583	208 (35.7%)	73 (35.1%)	Median = 62 months

FACS ⁴⁵ (2014)	2003- 2009	901	165 (18.3%)	68 (41.2%)	Median = 52.3 months
------------------------------	---------------	-----	-------------	------------	----------------------

Index years are the years during which patients were enrolled with index primary CRCs. Missing statistics were not available. Survival is broken out by location of disease when it was not available in a combined format.

Several other details of the FACS trial results corroborate the hypothesis that salvage surgery resulted in a survival benefit on average for treated patients. In the intention-to-treat-analysis, the median survival after any recurrence was greater in the collapsed intensive arms (combining the CT & CEA arm, the CT-only arm, and the CEA-only arm) than the minimal arm, although not significantly so (27.3 vs 14.6 months; $P = 0.11$). However, in the per-protocol analysis, this result was borderline significant ($P = 0.051$), and the same comparison for the two CT arms vs the two no-CT arms was significant ($P = 0.039$). The per-protocol analysis is potentially informative in this case because there was a non-trivial degree of contamination, mostly in the form of those in the minimal follow-up arm, and to a lesser extent the CEA-only arm, receiving additional CT scans. Finally, the ratio of the number of cancer-specific deaths to the number of recurrences was significantly lower in the three intensive arms than in the minimal follow-up arm ($P = 0.003$), suggesting that a smaller proportion of patients suffering recurrence died in the intensive arms.

While the benefit of aggressive salvage-surgery has not been established in an experimental setting, the above considerations corroborate the usage of salvage surgery when patients present with resectable, isolated disease. However, given that we might expect only 2-12% of a modern cohort of unselected patients to benefit from intensive-surveillance-induced salvage surgery (15%-30% recurrence rate and 15-40 percentage-

point increase in the salvage rate), it's unclear what kind of mortality benefit we could expect to see when averaged over the entire cohort. Moreover, if there were only a small benefit associated with intensive follow-up, it may not be surprising that recent trials failed to detect any mortality reduction.

In this paper, I use a modeling analysis to examine these issues. The analyses are based on the microsimulation model of CRC, recurrence-detection, treatment, and mortality described in Chapter 2. This model embodies the hypothesis that more aggressive extra-colonic follow-up of patients can increase the proportion of recurrences that are amenable to curative resection and thereby improve the survival of such patients. However, as described in Chapter 2, the model is also empirically grounded in the sense that the increase in salvage rates attributable to intensive follow-up in the model are consistent with meta-analytic results (Table 2) and the survival experience of patients treated with (without) salvage surgery is consistent with that observed in unselected cohort studies – Table 7 - where patients were treated aggressively (randomized control trials of systemic chemotherapy¹²¹) and, importantly, the FACS trial. Thus, a priori, the model, and therefore the underlying hypothesis it embodies, is theoretically consistent with many details of the FACS trial results. What remains to be determined is whether the primary outcome of the FACS trial – an all-cause mortality HR of 1.15 (95% CI: 0.87, 1.50) comparing a collapsed intensive arm to the minimal arm – and of the GILDA trial (HR = 1.14; 95% CI: 0.87-1.48) - is consistent with the causal theory embodied by the model. That is, the question is whether the primary results of these trials could be plausibly reproduced in a world characterized by the model.

In what follows, I demonstrate that the underlying causal theory of the model is in fact consistent with the results of the FACS and GILDA trials and that this realization should influence our interpretation of them and prompt reappraisal of recent criticisms of intensive follow-up. I first show that given the current low recurrence rates and the incremental salvage rates associated with intensive follow-up, any benefit is likely to be small and would require an impractically large trial to detect. To do this, I use Monte Carlo analyses to estimate what magnitude incremental benefit we would expect from the interventions under study in the trials if I am right that the benefit of intensive follow-up is limited by the low number of potential beneficiaries rather than the inefficacy of salvage surgery. I then use this result to perform a sample size calculation for a future hypothetical trial trying to detect the same effect. Secondly, I show that, in the case of the FACS trial, what was already a terribly underpowered study was further derailed by a sizeable imbalance in the recurrence rates between the arms. While some level of imbalance in the detection of recurrences may be expected given the asymmetrical follow-up intensities, I show that the observed level of recurrence imbalance is implausibly attributed to earlier diagnosis of recurrence in the intensive arms. In particular, I show that even after accounting for asymmetrical follow-up intensities a residual imbalance remains. I also quantify the expected bias and loss of power resulting from this recurrence imbalance. Finally, I perform a formal goodness-of-fit test comparing the model to the results of the FACS trial.

METHODS

The analyses of this paper involve the microsimulation model described in Chapter 2 and include (1) generating Monte Carlo estimates of the model-implied efficacy of the

surveillance regimens evaluated in the FACS trial and performing a sample size calculation for a hypothetical future trial, (2) an assessment of the degree to which the observed recurrence imbalance in the FACS trial could be explained away by earlier detection of recurrence due to more intensive follow-up, (3) performing a power analysis of the FACS trial given the observed recurrence levels but ignoring the recurrence imbalance, (4) an assessment of the expected bias and loss of power in the FACS trial from the chance recurrence imbalance, and (5), in light of the aforementioned power issues, a formal evaluation of the consistency of the model (and its underlying hypothesis) with the results of the FACS trial with a goodness-of-fit-test.

Estimation of Efficacy Parameters

To clarify what magnitude mortality reduction we could expect from intensive surveillance if the benefit arose exclusively from the survival advantage conferred by curative resection of recurrent disease, I conducted a Monte Carlo analysis comparing two of the surveillance regimens under study in the FACS trial. In particular, I compared the most intensive schedule – the use of routine CEA testing and CT studies - to the minimal regimen of 1 CT scan. The surveillance regimens for all 4 arms of the trial are shown in Table 8. I estimated location-, stage-, and adjuvant-therapy-specific all-cause and disease-specific mortality HRs at 5 years using US life-tables for background mortality. Monte Carlo precision was quantified with 95% confidence intervals. For all analyses, I simulated a cohort of patients that was 70 years old (the median age at diagnosis) and was 60% male.⁵⁹

I also estimated the approximate sample size that would be required in a future hypothetical RCT comparing routine CT and CEA testing to minimal follow-up (as

depicted in Table 8) in order for the trial to have an 80% chance of detecting the model-estimated all-cause mortality hazard reduction at 5 years. I assumed a type-I error of 0.05 and, for convenience, that there would be no loss to follow-up. I further assumed the trial would enroll stage II and III colon and rectal cancer patients only and after any adjuvant therapy (so beginning at 6 months). For stage III patients, I assumed 95% would receive adjuvant therapy (85% FOLFOX and 10% 5-FU/LV). While in a population setting about one-third of stage III patients receive no adjuvant therapy^{135 136}, it is likely that this figure would be significantly higher among patients who would enroll in a randomized control trial. For stage II patients, I assumed an equal proportion of patients would use FOLFOX, 5-FU/LV, and no adjuvant since the survival benefit of adding oxaliplatin to 5-FU/LV in stage II colon cancer patients is more controversial.¹⁹ I further assumed that 60% of enrollees would be male and that ages would vary from 40 to 78 with mean age of 68. I used this age distribution to mimic the likely age-distribution of a future trial (based on past trials). Finally, I estimated the statistical power of an alternative hypothetical trial comparing routine CT and CEA testing to the CEA-only (+1 CT scan) protocol with the above calculated sample size.

Table 8: Surveillance Schedule for the 4 Arms of the FACS Trial

Follow-up Arm	Test	3	6	9	12	15	18	21	24	30	36	42	48	54	60
CT + CEA	CT		X		X		X		X		X		X		X
CT-Only	CT		X		X		X		X		X		X		X
CEA-Only	CT					X									
Minimal	CT					X									
CT + CEA	CEA	X	X	X	X	X	X	X	X	X	X	X	X	X	X
CT-Only	CEA														
CEA-Only	CEA	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Minimal	CEA														

Table 8 depicts the follow-up schedule for each of the four arms in the FACS trial. The ‘CT + CEA’, ‘CT-Only’, and ‘CEA-Only’ arms are the intensive arms.

The above analyses are meant to provide detailed information about how big the benefit of an intensive follow-up regimen like that studied in the FACS trial might be in a US clinical practice setting, how the benefit might vary with disease location (rectal vs colon) and stage and adjuvant treatment, and what sized RCT would be required to detect such a benefit. In order to facilitate direct comparison of the model output with the results of the FACS trial, I conducted a second set of Monte Carlo analyses that mimicked the mix of patient and disease characteristics of the FACS trial. In particular, I estimated an all-cause and cancer-specific mortality HR at 5 years comparing (a) routine CT and CEA testing to minimal follow-up and (b) routine CT and CEA testing to CEA-only (+ 1 CT scan) testing (again as shown in Table 8). However, in this set of analyses, cohorts were simulated with the same mix of stage (I-III), adjuvant therapy utilization, and location (rectal vs colon) as reported in the FACS trial. Moreover, to better recreate the actual trial design, the life-course of all patients receiving adjuvant therapy was simulated for 5.5 years, but only the last 5 years were included in the analysis. As in the real trial, any of these patients who presented with recurrence within 6 months of the simulation were excluded from the analysis. For background mortality, I initially used lifetables from the UK for the period under study taken from the Human Life Table Database. However, observed mortality in the trial unrelated to CRC was much lower than would be predicted for the general population. I therefore injected mortality from other causes into the simulation by randomly assigning death times (from other causes) to individuals from an exponential distribution calibrated to match the cumulative incidence of non-cancer-

related mortality during the trial period (6.6%). Finally, since the primary outcome of the FACS trial was actually an all-cause mortality HR comparing the 3 intensive arms (collapsed into one arm) with the minimal arm, I replicated the above analyses for this comparison. Thus, one-third of patients in the intensive arm received intensive CEA and CT, one-third received intensive CT only, and the other third received intensive CEA and one CT scan. I also estimated the difference in five-year overall-survival probabilities between these two cohorts.

For the GILDA trial, I performed a similar analysis, but the capacity of the model to replicate the study design was somewhat limited. In its current form, the model does not distinguish recurrence by location, e.g., liver, lungs, local, peritoneum, etc, and thus it is difficult to mimic the protocol of the GILDA trial which involved the use of abdominal ultrasounds and chest x-rays at staggered intervals for colon cancer patients. For surveillance, the model currently can mimic the performance of CEA testing and an abdominal/thoracic (abdominal/thoracic/pelvic for rectal cancer) CT scan. As noted in Chapter 2, there is little chance of pulmonary recurrence in the first 12 months after primary resection, and so abdominal ultrasound without chest x-ray would reasonably be approximated by a CT scan in the simulation model. I thus simulated cohorts with a follow-up regimen that I thought would best represent the actual schedules offered in the trial given the constraints of the model. Table 9 displays the follow-up schedules of the actual trial and those used in the simulation study. On average, participants in the trial faced an annual cumulative probability of dying from other causes of about 0.5%, so I randomly selected 5 of every 1,000 participants to die each year.

Table 9: Surveillance Schedule for 2 GILDA Trial Arms

Follow-up Arm	Test	4	8	12	16	20	24	30	36	42	48	60
Intensive Arm: Real	Ultrasound/CT*	X	X	X	X		X		X		X	X
Intensive Arm: Real	Chest X-ray			X			X		X		X	X
Intensive Arm: Simulation	CT	X	X	X	X		X		X		X	X
Minimal Arm: Real	Ultrasound/CT		X		X							
Minimal Arm: Real	Chest X-ray			X								
Minimal Arm: Simulation	CT		X		X							
Intensive Arm	CEA	X	X	X	X	X	X	X	X	X	X	X
Minimal Arm	CEA	X	X	X	X	X	X	X	X	X	X	X

Table 9 lists the extracolonic-focused components of the 2 follow-up regimens under study in the GILDA trial. Colonoscopy and Proctoscopy are not listed.

*The GILDA trial protocol called for substituting some ultrasounds with abdominal/pelvic CT scans for rectal patients in the experimental arm and allowed 1 abdominal/pelvic CT scan for rectal patients in the control arm.

An Evaluation of Recurrence Imbalances

Over the course of the FACS trial, 12.6% of patients randomized to the minimal arm were diagnosed with any type of recurrence, and 18.3% of patients in the three more intensive arms suffered a recurrence. When time on study and censoring were accounted for, the respective 5-year net cumulative incidences of recurrence were 14.1% and 19.4%.¹¹ Among the 3 intensive arms, the routine CEA testing and CT scan arm

¹¹ I calculated these cumulative incidence values from curves using the previously mentioned PlogDigitizer software.

exhibited a 17.1% cumulative incidence while 20.5% of the CT-only and CEA-only (+ 1 CT) arms developed recurrence. These represent a 21% and 45% increase in the observed recurrence rate relative to the minimal arm, respectively.

To investigate whether this recurrence asymmetry could be explained by a more intensive follow-up regimen simply detecting recurrences earlier, I performed the following analyses. I first simulated the 5-year experience of a cohort of patients mimicking the disease and adjuvant-treatment characteristics of the patients in the FACS trial as well as the design features described above. Individuals were assigned in equal proportion to one of the three intensive follow-up schedules in the trial. Moreover, each individual's chance of suffering a recurrence (over the course of 15 years)¹² was adjusted by a calibrated factor such that, after 5 years, the cumulative incidence of recurrence in the absence of competing risks¹³ produced by the model matched that observed in the collapsed intensive follow-up arm of the FACS trial (19.4%). I then re-simulated the same cohort except this time they were subject to the minimal surveillance schedule (Table 8). This allowed estimation of the five-year cumulative incidence of recurrence that would be expected given the same underlying propensity for disease but only minimal follow-up. That is, this analysis served to quantify the difference in observed recurrence rates we would expect between the 3 intensive arms and the minimal arm in virtue of different follow-up intensities. Any further difference (over and above that)

¹² See Chapter 2 for details of the model. Recall that recurrence was simulated as a binary process (recur within 15 years or not) and then, conditional on eventually recurring, a time to recurrent disease process was simulated.

¹³ The cumulative incidence rates reported in the trial were estimated using Kaplan Meier methods that censored patients at the time of death from other causes.

would then represent a recurrence imbalance between arms. Finally, I repeated this set of analyses using just the routine CEA and CT arm as the intervention arm. This was done to evaluate whether this specific treatment-arm (which appeared to suffer less of a recurrence imbalance) would allow a less-biased comparison with the minimal-arm.

The GILDA trial exhibited a small recurrence asymmetry at 5 years as well. This was also in favor of the control arm (18.8% vs 22.0% in the intensive arm). I therefore repeated the above simulation analysis for the GILDA trial.

Power, Bias, and Goodness-of-Fit Analyses

Based upon the results of the above Monte Carlo analyses, I performed a simulation study to estimate the statistical power of the FACS trial to detect a mortality benefit of the size projected by the model. Since the primary outcome of the study was the all-cause mortality HR comparing the collapsed intensive arm (N = 901) to the control arm (N = 301), I conducted the power analysis for this comparison. I performed the simulation study twice, and in each study I simulated 2000 replications of the trial. In both cases, I simulated the protocol follow-up schedule and not the observed degree of follow-up. That is, lack of compliance and contamination were ignored for these analyses.

In the first simulation study, in both arms, I adjusted each individual's risk of recurrence by the factor that was previously calibrated to reproduce the trial-observed 5-year cumulative incidence of recurrence (19.4%) in the collapsed intensive arm. This means that simulated patients in the minimal arm were adjusted by the same factor and so had the same propensity for recurrence as those in the collapsed intensive arm. Thus the first set of analyses ignored any underlying recurrence imbalance. I further used the results of the first simulation study to construct a sample distribution of the all-cause

mortality HR estimator and the observed 5-year cumulative incidence of recurrence in the minimal arm. I calculated the 0.025 and 0.975 quantiles of each distribution. I also determined the empirical quantile (i.e., quantile of the sample distribution) of the actual FACS-trial-observed all-cause mortality HR estimate and minimal-arm recurrence rate.

The second power-analysis differed only in the adjustment to the risk of recurrence. In this simulation study, I specified the number of patients recurring in both the minimal arm and the collapsed intensive arm. These numbers were calibrated to ensure that the collapsed intensive arm (minimal arm) exhibited a 5-year cumulative incidence of 19.4% (14.1%) in the absence of background mortality and using the FACS design. In order to select which patients in each arm suffered a recurrence, I (pseudo)randomly sampled from a discrete distribution where each person's chance of being selected was equal to their risk of recurrence renormalized so that all such probabilities summed to 1. Thus, in each of the 2000 simulations of the trial, the arm-specific recurrence rates at 5 years matched those observed in the FACS trial.¹⁴ Moreover, any underlying recurrence imbalance backed out in the previous section (over and above what would be expected due to asymmetrical follow-up intensity) was present in every simulation.

The first simulation study serves to estimate the statistical power a trial like the FACS trial would have to detect a mortality reduction the size of the model-predicted all-cause mortality HR. It is based on the sample distribution of the HR estimator assuming the model-predicted HR is the true parameter and that both arms had the same underlying

¹⁴ Strictly speaking, there was a small amount of variation due to variation in when a few patients' disease was detected (within 5 years or not). However, this was no more than +/- 0.4 %.

propensity for disease. Though, given the small number of recurrences, the actual observed recurrence rates in each arm will vary from sample to sample.

The second simulation study estimates the power the FACS trial had to detect the same magnitude mortality reduction except now conditional on the backed-out arm-specific recurrence rates so that the actual 5-year cumulative incidence of diagnosed disease in each arm matched that observed in the trial. It is thus based on the sample distribution of the HR estimator conditional on the model and the actual underlying recurrence imbalance. The difference between the results of the two simulation studies quantifies the loss of statistical power due to the observed underlying recurrence imbalance, conditional on the model-predicted true efficacy.

Using the second simulation study described above, I also estimated the bias induced by the chance recurrence imbalance and performed a goodness-of-fit test comparing the model with the observed FACS trial results. By bias, I mean the difference between the large-sample (or true) model-implied HR and the average HR estimated across repeated trial simulations, conditional on any underlying recurrence imbalance. I estimated both the absolute bias and the relative bias, and I calculate both on the log scale. The latter is defined as the absolute bias divided by the true parameter. Finally, using the resulting sample distribution for the all-cause mortality HR estimator (conditional on any underlying recurrence imbalance), I calculate the probability of observing a result at least as extreme as the actual HR estimate reported for the trial. This probability is analogous to a frequentist two-sided p-value. Following the logic of predictive model checking, the point is to formally assess whether, given the recurrence imbalance, there are “systematic differences between the model and some aspects of the data”.¹³⁷

RESULTS

Estimation of Efficacy Parameters

Table 10 contains the location-, stage-, and adjuvant-therapy-specific simulation results. The fourth column of Table 10 lists the average (across parameter sets) proportion of recurrences presenting within 5 years for which R0 salvage was possible in the intensive (CT + CEA) arm. These varied from 0.37 for stage III colon cancer with no adjuvant treatment to 0.47 for stage II colon cancer with FOLFOX. Disease-specific mortality HRs comparing the intensive arm to the minimal arm varied from about 0.89 (stage II colon no adjuvant) to 0.95 (stage II colon FOLFOX). These were influenced by the difference between the R0 salvage rate in the intensive arm and the minimal arm, the stage of the patients (the expected benefit of salvage is greater on average for stage I/II patients), and most importantly the proportion of patients who suffered a recurrence. All-cause mortality HRs at 5 years ranged from about 0.93 (stage III colon with no adjuvant treatment) to 0.99 (stage I colon). These were influenced by the disease-specific hazard ratio and the ratio of noise (background mortality) to cancer-specific mortality.

For a hypothetical future trial comparing the CT + CEA follow-up regimen to minimal follow-up (as depicted in Table 8) and enrolling only stage II and III colon and rectal patients, we could expect about 21% of patients to suffer a recurrence within 5 years and the intensive arm to procure about a 25 percentage-point increase in the R0 salvage rate. This translates into about a 9% reduction in disease-specific mortality and, assuming background mortality rates equivalent to US lifetables, a 5% reduction in overall mortality (Table 10). I estimated that in order for such a trial to have an 80% chance of detecting that effect ($p < 0.05$) it would need about 18,000 (+/- 1,000)

participants per arm. Thus even assuming there would be no loss to follow-up, the trial would need at least 34,000 enrollees. For the comparison of the CT + CEA to CEA-only (+ 1 CT scan) follow-up schedules, we could expect about a 6 percentage-point increase in the R0 salvage rate and only about a 1% reduction in disease-specific and overall mortality hazards at 5 years. At a sample size of 18,000 participants per arm, I estimated such a trial would have less than 1% power to detect an effect of such a small magnitude.

The model was able to replicate the results of the FACS and GILDA trials reasonably well. In general, the model-estimated R0 salvage proportions for each of the four follow-up regimens of the FAC trial depicted in Table 8 were within a few percentage points of the actual FACS-observed rates (Table 10). The only exception to this was the CT-only arm. The model-estimated R0 salvage rate – 38% - was quite a bit lower than the rate observed in the FACS trial - 46% (95% CI: 34%-58%). However, the trial estimates were very imprecise and the model-estimate was within the range of the 95% confidence interval. The model-estimated minimal-arm R0 salvage-rate of 17% was closer to the rate reported in the per-protocol analysis (14%) of the FACS trial than to the rate reported for the intention-to-treat analysis (21%). The latter was influenced by non-trivial contamination in the form of additional CT-scans. As can be seen in Table 10, the model-predicted true effects of the interventions under study were small. We could expect a 6% all-cause mortality-hazard reduction in upgrading from the minimal follow-up arm (1 CT scan) to the most intensive arm (CT + CEA). For the comparison involving the collapsed intensive-arm, this figure was 5%. Moreover, the chance of patients in the collapsed intensive arm surviving to 5 years was approximately 1 percentage point greater than that

of patients in the minimal arm. For the GILDA trial, we could only expect about a 1-2% mortality hazard reduction.

Table 10: Cohort-Specific All-Cause and Disease-Specific Mortality Hazard Ratios at 5 Years

Cohort	All-Cause Mortality HR (95% Monte Carlo CI)	Disease-Specific Mortality* HR (95% Monte Carlo CI)	Model-Estimated (Trial-Observed) Intensive-Arm Salvage Rate#
<i>Disease-Specific Simulation Results Comparing CT + CEA vs Minimal Follow-up Using US Lifetables^{##}</i>			
Stage III Colon (FOLFOX)	0.951 (0.947, 0.955)	0.927 (0.922, 0.932)	0.39
Stage III Colon (FU + LV)	0.935 (0.932, 0.940)	0.910 (0.906, 0.915)	0.38
Stage III Colon (No Adjuvant)	0.926 (0.922, 0.930)	0.909 (0.905, 0.913)	0.37
Stage III Rectal	0.946 (0.941, 0.952)	0.923 (0.918, 0.927)	0.41
Stage II Colon (FOLFOX)	0.980 (0.975, 0.986)	0.953 (0.943, 0.963)	0.47
Stage II Colon (FU + LV)	0.957 (0.952, 0.963)	0.904 (0.896, 0.911)	0.42
Stage II Colon (No Adjuvant)	0.938 (0.932, 0.944)	0.886 (0.878, 0.893)	0.39
Stage II Rectal	0.941 (0.936, 0.946)	0.897 (0.889, 0.905)	0.43
Stage I Colon	0.986 (0.980, 0.992)	0.941 (0.929, 0.953)	0.46
Stage I Rectal	0.973 (0.967, 0.980)	0.910 (0.903, 0.917)	0.44
<i>Model-Predicted True Effects for a Hypothetical Future Trial Using US Lifetables[^]</i>			
Future Trial: CT + CEA vs Minimal	0.948 (0.945, 0.951)	0.912 (0.909, 0.916)	CT+CEA = 0.41 Minimal = 0.16
Future Trial: CT + CEA vs CEA	0.988 (0.986, 0.991)	0.978 (0.975, 0.982)	CT+CEA = 0.41 CEA = 0.35
<i>Model-Predicted True Effects for FACS & GILDA Trial Simulations^{**}</i>			
FACS Trial (CT + CEA vs Minimal)	0.940 (0.936, 0.944)	0.912 (0.907, 0.917)	CT+CEA = 0.41 (0.44; 95% CI: 0.31-0.58) Minimal = 0.17 (0.21; 95% CI: 0.11-0.36)
FACS Trial (CT + CEA vs CEA)	0.992 (0.989, 0.996)	0.986 (0.981, 0.992)	CT+CEA = 0.41 (0.44; 95% CI: 0.31-0.58) CEA = 0.35

			(0.34; 95% CI: 0.23-0.47)
FACS Trial: Collapsed Intensive vs Minimal	0.948 (0.945, 0.951)	0.923 (0.921, 0.926)	Intensive = 0.38 (0.41; 95% CI: 0.34-0.49) Minimal = 0.17 (0.21; 95% CI: 0.11-0.36)
GILDA TRIAL	0.987 (0.982, 0.991)	0.981 (0.978, 0.985)	Intensive = 0.42 (0.42; 95% CI: 0.34-0.51) Minimal = 0.35 (0.40; 95% CI: 0.32-0.49)

* The disease-specific mortality hazard ratio was calculated in the absence of competing risks.

The R0 curative rate in the intensive arm averaged across different parameter sets.

Patients in each cohort were 70 years old and 60% were male.

^ Cohorts were split evenly between stage II and III disease. For Stage III patients 85%, 15%, and 5% were assumed to have received FOLFOX, 5-FU/LV, and surgery-only, respectively. These figures were one-third each for stage II patients. Sixty-percent of patients were male and ages ranged from 40-78 with a mean of 68.

**Follow-up regimens for the FACS (GILDA) trial are given in Table 8 (Table 9). Demographic disease characteristics mimicked those in the trial.

An Evaluation of Recurrence Imbalances

The model-simulated collapsed intensive arm of the FACS trial had a 5-year cumulative incidence of recurrence of 19.2%. This was very close to the 19.4% observed in the actual trial. The factor of adjustment was thus only trivially different from 1. The corresponding model-simulated 5-year recurrence rate in the minimal arm was 17.7% (compared with 14.1% in the real trial). Thus, while we would expect about a 9% reduction in the observed recurrence rate in the minimal arm compared to the intensive arms in virtue of patients only undergoing 1 CT scan, we observed a 27% reduction. In absolute terms, the respective figures are a 1.7 percentage-point reduction and a 5.3 percentage-point reduction. For the CT + CEA vs CEA only (+ 1 CT scan) comparison in

the FACS trial, there was less of an unexplained recurrence imbalance. After the model-simulated 5-year cumulative incidence of recurrence in the CT + CEA arm was adjusted down to 17.1% (from roughly 20%), the model-simulated minimal arm exhibited a 5-year recurrence rate of 15.3% (as opposed to 14.1%). For the GILDA trial, the model-simulated recurrence rate in the intensive arm was 21.4%. This was minimally readjusted to get a recurrence rate of 22.0% (that actually observed in the intensive arm of the trial). The associated recurrence rate in the control arm predicted by the model was 21.2%. The actual observed rate in the GILDA trial was 18.8%. Thus, in all 3 cases the trial control arms exhibited a greater drop in recurrence rates compared to the experimental arm than would be expected in virtue of asymmetrical follow-up intensity alone.

Power, Bias, and Goodness-of-Fit Analyses

Given the small size of the model-estimated mortality hazard reduction associated with the FACS intensive arms, the results of the power calculations are unsurprisingly discouraging. Given the recurrence rates exhibited in the 3 intensive arms but ignoring the recurrence imbalance, the FACS trial likely had between 1-3% power to detect the 5% all-cause mortality reduction. The actual point estimate reported in the trial (HR = 1.15) was barely consistent with the model-estimated sample distribution of the HR estimator: 2.5th and 97.5th percentiles equal to 0.74 and 1.15, respectively. However, this is less problematic for the causal theory embodied in the model when we consider that the observed recurrence rate in the minimal arm of the FACS trial was also extreme. The model-estimated 2.5th and 97.5th percentiles of the sample distribution of the minimal-arm recurrence rate were 13.7% and 22.0%, respectively. The observed rate of 14.1% represents the 3.1 percentile of the distribution. However, the observed recurrence rate in

the collapsed intensive arm (19.4%) represents the 54th percentile of its corresponding sample distribution. It is worth restating that these results assume that participants in different arms had an identical underlying propensity for disease (namely that observed in the intensive arms) as would be expected across repeated samples of a well-designed RCT. The variation in recurrence rates from sample to sample is simply the result of randomness and of a small number of patients suffering a recurrence.

Conditional on the observed chance recurrence imbalance¹⁵, the FACS trial had essentially 0% power (95% CI: 0%, 0.7%) to detect the 5% mortality reduction. Though, I showed above that it would have also had close to 0% power in the absence of the recurrence imbalance. Moreover, the log-HR estimator was biased upwards by 0.19 (95% Monte Carlo CI: 0.18-0.21) due to the low recurrence rate in the control arm. The magnitude of this bias is 3-4 times the size of the model-estimated true effect (-0.05). On the HR scale, the trial-reported point estimate of a 15% increase in the mortality hazard in the intensive arm was identical to the model-predicted expected value of the estimator conditional on the recurrence imbalance: $E[\widehat{HR}] = 1.15$ (95% Monte Carlo CI: 1.14-1.17). There was thus no evidence of inconsistency between the predictions of the simulation model and the results of the trial ($P = 0.95$).

DISCUSSION

¹⁵Strictly speaking, by ‘conditional on the observed chance recurrence imbalance’, I mean conditional on the discrepancy between arms in terms of the underlying propensity for disease that was backed out so as to lead to the observed recurrence imbalance (taking into account the arms faced asymmetrical follow-up intensities)

The above results illustrates three main points. The first is that, if the benefit of intensive surveillance comes from earlier detection of recurrence leading to increased chances for curative resection, the benefit is small. Depending on the disease stage and location and the adjuvant chemotherapy received, providing patients with routine CT scanning and CEA testing can be expected to reduce overall-mortality hazards by about 1-8 percent compared to offering just 1 CT scan at 12-18 months. In the case of the FACS and GILDA trials, the hypothesized true benefits of the interventions under study were roughly a 5 and 1 percent reduction in mortality hazards at 5 years, respectively.

The second point is that although the FACS and GILDA trials were the largest and most relevant trials they had no chance to detect the above hypothesized effects. The authors of the report on the FACS trial estimated that they had a 31% chance of detecting a 5 percentage-point increase in the survival proportion at 5 years. However, the model-predicted survival risk-difference at 5 years was only 0.01 (rather than 0.05), and the trial had less than a 2% chance of detecting that effect. While I did not perform a formal power analysis on the GILDA trial, it is obvious that the chance of detecting a 1-2% mortality hazard reduction at 5 years with a trial of roughly 1,230 patients is trivial.

The third point is that it is very unlikely that a trial with adequate power to detect such a small hypothesized effect would ever or could ever be performed. There are roughly 150,000 new cases of CRC per year in the US. In order to enroll 36,000 patients over the course of 5 years, this would require enrolling approximately 5% of new cases per year. In this context, it is worth noting that only about 3% (1.5%) of US adult cancer patients under 65 (≥ 65) years old participant in any randomized control trials, let alone a single trial.¹³⁸ In reality, such a trial would likely require coordination across North

America, Europe, and likely Asia. It is worth noting that the FACS (confined to the UK) trial had originally planned to enroll a sufficient number of patients to have 80% power to detect a 5 percentage-point risk difference in survival at 5 years but, because of poor enrollment, was forced to abandon that goal and change its primary endpoint to a risk difference of R0 salvage rates. Similarly, the GILDA trial had originally planned to enroll around 4,000 participants with the hopes of detecting an overall mortality reduction of 20% (HR = 0.8) at 5 years but also had to drop that goal due to poor enrollment. The hypothetical trial considered here would require roughly 6-10 times as many patients. Finally, it is also unclear if such a logistically complex and inevitably expensive trial would even be worth the cost and thus whether it should ever be performed.

Given contemporary recurrence rates, it is unlikely that an RCT of intensive follow-up will ever settle this issue. The most efficient and probably only realistic way to determine if intensive follow-up could even offer the small benefit hypothesized to be associated with an increased rate of curative resection would be to investigate the efficacy of salvage surgery directly. The idea would be to aggressively follow-up a large cohort of CRC patients after curative resection of primary disease and then randomize all patients identified as having 'resectable' recurrent disease to attempted curative resection or conservative treatment. If the trial failed to show a benefit for salvage surgery, we could seriously consider throwing in the towel on intensive follow-up. On the other hand, a documented survival advantage from surgical resection with a curative intent would corroborate the thesis of this paper and support the use of intensive follow-up. Though the question of what follow-up regimen to use would remain an open question. A

feasibility study is currently underway in the UK to assess whether it will be possible to conduct a RCT to assess the efficacy of pulmonary metastectomy.¹³⁹

My results suggest that there was indeed a chance recurrence imbalance between the collapsed intensive arm and the minimal follow-up arm of the FACS trial. Given the main two points made above (a small benefit and essentially no power), the recurrence imbalance is really an issue of secondary importance. However, it is relevant to one possible criticism. Most of the trials in earlier meta-analyses found a non-significant mortality reduction from intensive follow-up compared to minimal or no follow-up. It might be expected that we should see a similar small benefit, even if nonsignificant, in the FACS and GILDA trials, but instead we saw the opposite ($HR > 1$). By examining the likely effect of the recurrence imbalance, I have shown that the results we saw in the FACS trial are exactly what the model would have predicted in the presence of such a recurrence imbalance.

The recurrence imbalance was notably smaller when comparing the routine CT and CEA arm to the minimal follow-up arm. It is likely that a direct comparison of these two arms would represent a less biased estimate of the true effect. Unfortunately, this result is not reported in any of the reports of the FACS trial. When I used methods described in Parmar et al., (1998)⁴² and Tierney et al., (2007)⁴³ to estimate the HR based upon digitized survival curves, I estimated an all-cause mortality HR of 0.94 (95% CI: 0.63-1.14) comparing the routine CT and CEA arm to the minimal arm. Unfortunately, the quality of the published figure and thus these results is unclear. However, if approximately correct this point estimate would nearly match the model-estimated true effect for this comparison.

I did not conduct the same formal analysis for the GILDA trial. However, given the model-predicted true effect of only a 1.3% mortality hazard reduction at 5 years and the demonstrated presence of a small recurrence imbalance that could not be attributed to asymmetrical follow-up intensities, the trial result of a nonsignificant 15% increase in mortality in the intensive arm is not surprising.

It is unclear whether the recurrence imbalance in the FACS trial arose because of an imbalance in risk factors for recurrence or because of random variation and a small recurrence size. My analyses suggest there would be roughly a 6 percent chance of the minimal arm exhibiting a recurrence rate at least as extreme (high or low) as that observed in the real trial if both it and the collapsed intensive arm had the same underlying risk of recurrence. The randomization scheme involved a minimization algorithm to increase the chances of balance among patient age, sex, and adjuvant treatment. A relatively good balance was achieved with regards to location of tumor (rectum, left colon, or right colon) and stage (I-III). However, it is possible there could have been an imbalance in other risk factors such as tumor grade or N and T status within stage (e.g., N1 vs N2 or T3 vs T4). Anyhow, regardless of the true cause, the effect was the same.

My analyses demonstrated that the observed recurrence imbalance was larger than any anticipated imbalance due to different intensity of follow-up and that the size of this residual imbalance was enough to dominate the model-predicted true effect size. The fact that the CT-only and CEA-only (+ 1 CT scan) arms had a cumulative incidence of recurrence roughly 3 percentage points higher (20.5% vs 17.1%) than the most intensive arm (routine CT + CEA) corroborates this result. There is no reason to expect such an

asymmetry to arise from follow-up differences. Moreover, prior empirical work further supports the validity of this conclusion. Trials of intensive surveillance have generally spanned 5 years after primary surgery. As mentioned earlier, a large majority of recurrences of any type will present within 5 years, a statistic that has been observed outside of trials of intensive surveillance. This is particularly true among stage II and III patients, the very patients who will present with most of the recurrences. The initial report on the FACS trial results was published prior to a median of 5 years of follow-up⁴⁶, but the recent manuscript summarizing the mature results (which were used in my analyses) reported a slightly amplified imbalance after a median (minimum) follow-up of 8.7 (5) years.⁴⁵ It is thus very unlikely that there would be many recurrences in either arm if follow-up continued. Aside from the smaller imbalance observed in the GILDA trial, the only other trial which demonstrated a non-trivial recurrence imbalance had the opposite outcome – more recurrences in the control arm.⁴¹ A more plausible explanation of this phenomenon is that stratification of the randomization procedure on stage (I-III, as opposed to say the American Joint Committee on Cancer TNM system) and/or adjuvant chemotherapy is insufficient to reduce the chances of a non-trivial recurrence imbalance across arms to an acceptable level in the presence of a low risk of recurrence and small trial sizes.

Still, a skeptic may suggest that perhaps there is a fundamental difference between the earlier-detected and later-detected recurrences in that the latter are far less aggressive and possibly in some cases not fatal. Perhaps some of these cases never would have even presented as clinical disease in a typical patients' remaining lifetime. That is, perhaps

these extra recurrences represent a form of over-diagnosis, similar to the much-discussed concept in a screening context.¹⁴⁰

However, this response is unsatisfactory. It is very unlikely that there is much use for the concept of over-diagnosis with respect to CRC recurrence. While there is evidence that patients with recurrence diagnosed after 3 years tend to have a slower-moving natural history⁵⁶, a large cohort study found no difference in time to death between patients diagnosed with recurrence between 2-5 years and those diagnosed after 5 years⁵⁴. Anyhow, since recurrence generally represents systemic disease, it is almost always fatal without curative resection. For example, approximately 75% of patients diagnosed with recurrence four years after treatment for the primary cancer will be dead within 5 years after recurrence.⁵⁶ It is also worth noting that the model heterogeneity parameters were calibrated so as to recreate the phenomenon of slower-moving natural history among patients diagnosed 3 years or after and my analyses still found the recurrence imbalance lead to a bias of a magnitude roughly 3-times larger than the model-predicted true effect size.

The argument of this paper is also relevant to the interpretation of two large RCTs evaluating intensive follow-up that are currently in progress. The Scandinavian COLOFOL trial¹⁴¹ (N = 2,500) has closed enrollment and is likely to report results very soon. The trial compares abdominal CT, chest x-ray, and CEA assay at 6, 12, 18, 24, and 36 months vs. at 12 and 36 months among stage II and III patients. The French FFCD PRODIGE 13 trial¹⁴² (N = 2,000) is nearing the end of its scheduled follow-up. Like the FACS trial it uses a 2x2 design, comparing two intensive imaging strategies (one with abdominal/pelvic/thoracic CT and one with just ultrasound) and the use of CEA vs no

CEA among stage II and III patients. It is extremely unlikely that either trial will detect any significant mortality benefit. If they did, it would suggest that intensive follow-up likely confers a survival advantage over and above that attributable to the increased rate of curative resection. For example, it was previously suggested¹⁴³ that intensive follow-up might improve overall survival via providing increased psychological support or through earlier detection and better management of other medical conditions. This would be particularly relevant for earlier trials in which patients in the control arm underwent no or very limited follow-up.

Another possible source of benefit from intensive follow-up would be the existence of a survival advantage among patients with unresectable and thus terminal disease from commencing systemic chemotherapy while the disease was still asymptomatic. There is weak evidence from a single randomized control trial to support the existence of such a benefit.¹⁴⁴ The trial (N = 183) randomized patients with advanced and unresectable yet asymptomatic CRC to immediate treatment with chemotherapy (5-FU/LV) or to a conservative approach that waited to start chemotherapeutic treatment until symptoms developed. The latter mimicked a counterfactual world in which the recurrent disease was not detected until symptomatic. There was a borderline significant survival benefit (increase of median survival by 5 months) among those randomized to immediate chemotherapy. The benefit appeared immediately and then diminished until it had completely disappeared by 12-18 months. A log-rank test failed to reject the null hypothesis after 3 years, but a Breslow-Gehan test was significant ($P < 0.02$). This latter hypothesis test prioritizes discrepancies between two survival curves that occur earlier on.¹²² There appears to be no further experimental evidence regarding such an effect,

perhaps for ethical reasons. Given the small size of the trial, the questionable significance, and the reality that treatment of terminal systemic disease has changed dramatically (with the development of new chemotherapeutic agents and targeted biological agents), it is unclear if such a benefit really exists or, if it once did, would still exist in contemporary practice.

The existence of such a benefit (roughly 5-months of extra life-expectancy from starting systemic chemotherapy before symptoms appear) along with the very high recurrence rates previously exhibited by CRC patients could help explain the more promising results of earlier trials and meta-analyses. That is, if the typical patient with recurrence received a 5-month survival advantage from asymptomatic detection (over and above any lead time) even if no curative resection was performed, and if 40-70% of patients suffered a recurrence, it would not be surprising if intensive follow-up lead to a notable mortality reduction. For example, the largest and most precise benefit found in the earlier trials was in Secco et al. (2002).⁴¹ The description of this trial was somewhat unclear, but it appears to have enrolled only rectal cancer patients. Moreover, the trial used stratified randomization based upon the risk of recurrence (high vs low risk), and, in the high-risk groups, roughly 70% of patients suffered a recurrence. In addition, high-risk patients in the experimental arm underwent intensive imaging while the same patients in the control arm had no scheduled follow-up. Unfortunately an all-cause mortality HR was not reported for either risk-stratified group or overall. However, using the previously mentioned methods^{42 43} and digitized survival curves, I estimated that high-risk participants of the experimental arm exhibited a nearly 40% reduction in the mortality hazards at 5 years (HR = 0.61; 95% CI: 0.42, 0.87).

This combination of factors may shed light on previous results. However, given contemporary recurrence rates, and in light of the results presented in this paper, it is unreasonable to expect to replicate anything close to the 25% mortality-hazard reduction (HR = 0.75; 95% CI: 0.66-0.86) that was reported in a 2015 meta-analysis.³⁷ The subsequent review by Jeffery et al., (2016)⁴⁷ reported a meta-analytic all-cause mortality HR of 0.90 (95% CI: 0.78-1.02). Although this estimate does not rule out the possibility of no benefit on average, this is consistent with the existence of a benefit of the magnitude suggested by my analyses. Moreover, given the extremely large sample size that would be needed in order for a single trial to have reasonable power to detect the hypothesized effect, even if such a benefit existed it would not be surprising that a meta-analytic interval estimate was unable to exclude no-effect on average.

The most recent review by Mohkles et al., (2017)⁴⁸ estimated an all-cause mortality meta-analytic HR of 0.98 (95% CI: 0.87-1.11). A thorough evaluation of the decisions of these reviewers is beyond the scope of this paper, but some of them can certainly be questioned. Firstly, they excluded the above-mentioned trial Secco et al., (2002). The reviewers are right that there was an unexplained imbalance in arm sizes. Though, the authors did refer to a randomization process. However, the reviewers oddly and incorrectly claim that the trial results did not appropriately address censoring but instead excluded all such patients. Anyhow, this trial is not the only trial with a potentially concerning risk of bias. The GILDA trial (which they included) suffered from a sizable attrition problem with about 27-29% of participants in each arm being lost to follow-up over the course of 5 years. This figure does not include patients who exited follow-up because they (a) developed a new and unrelated type of malignancy or (b) died from

other causes. In fact, no reasons are given for why patients were lost to follow-up. Though admittedly there was not any serious asymmetry between trial arms, and the analysts did use appropriate methods to adjust for time on study and censoring. However, there is no mention or discussion of use of death registries to monitor the vital status of lost patients, and so it is unclear if the investigators had access to participants' health records, disease-status, or vital-status.

Secondly, the random-effect model reported in Jeffery et al., assigned the CEASL trial (mentioned in the introduction) 50% of the weight even though it is unclear how relevant the trial is to contemporary practice. The 'intensive' arm of the CEASL trial bears little resemblance to any surveillance protocol that would be considered today. Unlike most of the other earlier trials, it included no imaging studies, e.g., abdominal US or chest x-ray, and no endoscopy. It is also unclear if surgeons based decisions on whether to operate or not on any type of diagnostic imaging to verify that recurrence was causing the elevated CEA levels. The suspicion that these were not performed is corroborated by the very high attempted salvage rate in the 'intensive' arm. Terminology is unclear, but it appears that 68% of the 108 patients randomized to the 'aggressive' arm underwent at least an exploratory 'minimal-laparotomy' and about 84% of these patients proceeded to a full laparotomy. In most of the other trials conducted around the same time (1980's), the treatment arm exhibited a salvage rate of under 30%. These other trials included routine endoscopy and imaging in their intensive arms and so would have likely had a better (i.e., earlier in the natural history process) risk-pool of patients to work with and thus higher salvage rates. Thus it seems likely that the 'aggressive arm' performed what would now be considered inappropriate surgery on a massive scale. At the time of

the CEASL trial, 30-day surgical mortality rates of over 5% were common.¹⁴⁵ This could certainly have watered down any potential benefits from attempted salvage surgery.

CONCLUSION

I have argued that it is misguided to interpret the negative results of the FACS and GILDA trials as evidence of no benefit of intensive surveillance and particularly as evidence of no benefit from aggressive surgical treatment of recurrence. While it is true that the trials failed to show evidence of benefit, the trials likely had no chance of detecting the hypothesized effect, and the results of the FACS trial (a median survival after curative resection equivalent to that reported in case-series) were consistent with the theory that a large portion of patients with recurrence can benefit from curative resection. These facts should be considered in future iterations of professional guidelines.

Although these conclusions depend on the validity of the microsimulation model described in Chapter 2, there is good reason to believe it adequately reflects the important characteristics of disease natural history and the process of disease detection and treatment to offer valuable insight into the clinical problem. As shown in Appendix B, the model was generally able to replicate the time-to-diagnosed-recurrence hazard rates well. It also mimicked two important characteristics of natural history that could influence an evaluation of surveillance: (1) patients who present with a shorter disease-free interval tend to have a quicker time to death in the absence of surgical cure and (2) patients with node-positive primary disease face a worse prognosis after an attempted curative resection. These matter because most recurrences occur earlier and a higher proportion of stage III patients suffer recurrence. The model was also able to recreate differences in the R0 salvage rate across disease stage (Table 10) - with more option for salvage among

stage I and II patients^{45 146} - even though I did not explicitly model this phenomenon of between-stage variation in salvageability. This suggest that this variation may arise from differences in the timing at which patients present with recurrent disease or differences in sojourn time rather than say stage III patients tending to exhibit more widely-disseminated disease. The model also generally was able to recreate the R0 salvage rates of the FACS and GILDA trial arms. Finally, the mortality experience of patients treated palliatively and with surgical cure mimicked that reported in randomized control trials and/or unselected cohort studies, respectively.

In the absence of a direct RCT of the efficacy of salvage surgery or an extremely large and probably impractical trial of intensive follow-up, future guidelines and clinical-practice decisions will have to be made with only weak and inconclusive evidence. Some will balk at continuing to recommend a clinical protocol for hundreds of thousands of patients without any clear experimental evidence. The risks here are wasted medical recourses, unnecessary morbidity and mortality associated with salvage surgery, and wasted patient time. However, there are risks associated with prematurely backing away from aggressive follow-up. If there is indeed the projected benefit represented by the microsimulation model, reducing the intensity of follow-up or scraping it altogether could deprive patients of life-prolonging and in some cases life-saving procedures.

Given the small size of the hypothesized benefit, it may be the case that intensive follow-up regimens are not cost-effective, particularly those involving intensive imaging. It is thus worth considering what the optimal schedule of CT imaging follow-up would be if the hypothesized benefit did in fact exist. Given the corroborating empirical evidence

in support of the efficacy of aggressive salvage surgery discussed above, following such a strategy might represent a reasonable compromise. In Chapter 4, I turn to this task.

CHAPTER 4: **A Cost-Effectiveness Analysis of Alternative Post-Diagnosis Surveillance Strategies for Colorectal Cancer**

INTRODUCTION

As explained in Chapter 1 of this dissertation, the stated goal of routine extra-colonic surveillance following curative resection for stage I-III CRC is to detect late-occurring metastatic disease or extramural local-regional recurrence while the disease is asymptomatic and hopefully sufficiently isolated to be amenable to curative resection, i.e., salvage treatment. As a reminder, multiple meta-analyses have pooled the results of different combinations of the 13 RCTs that have evaluated alternative intensities of surveillance, 11 of which evaluated more vs less intensive extra-colonic-focused modalities (as opposed to endoscopic surveillance). While earlier meta-analyses found intensive follow-up using CEA testing and radiological imaging reduced overall mortality, the two most recent analyses concluded that it failed to provide such a benefit. This in turn was partially driven by the selection of what studies to include but mostly by the recent publication of 3 trials – the FACS, GILDA, and CEASL trials - showing intensive surveillance was associated with a non-significant increase in the mortality hazard.

In Chapter 3 of this dissertation, I argued that there was evidence (albeit arguably weak and certainly non-experimental) for the theory that curative surgical resection could offer a survival advantage for patients who suffer a recurrence, even when performed at aggressive levels (e.g., 30-40% of patients who undergo routine surveillance). Moreover, I argued that, even though the FACS and GILDA trials were the largest and most recent,

it was a mistake to interpret their negative results as evidence of there being no benefit from such follow-up. I showed that if intensive extra-colonic surveillance did in fact confer a survival advantage in virtue of increasing the chances of clear-margin surgical resection of recurrent disease then the benefit would be extremely small when averaged over an entire cohort of patients undergoing such surveillance. As such, even ignoring the recurrence imbalance observed in the trials (particularly the FACS trials), the FACS and GILDA trials would have had only trivial power to detect such an effect. Moreover, I questioned the relevance of the CEASL trial and some of the inclusion/exclusion decisions made by the authors of the most recent meta-analyses and suggested that an alternative attempt at systematic review and meta-analysis would likely find a small (<10%) but non-significant reduction in all-cause mortality hazards on average. Finally, I argued that, unfortunately, it likely will not be logistically or financially feasible to enroll a sufficient number of patients to settle this issue with a new RCT evaluating intensive surveillance. An alternative possibility that some will no doubt find objectionable would be to conduct a large RCT directly evaluating surgical resection with curative intent by randomizing all patients with 'resectable disease' identified in a program of intensive surveillance to surgical resection or palliative care.

In the absence of such evidence, policymakers and panel members responsible for issuing professional-society guidelines face a difficult choice. Should they recommend patients undergo, and reimburse for, intensive extra-colonic surveillance even though such a practice is not supported by anything more than weak evidence? Or should they take a more conservative approach and balk at recommending or paying for medical care that is not grounded in experimental evidence. In this context, a cost-effectiveness

analysis could be fruitfully used to shed light on the decision-problem. In particular, a cost-effectiveness analysis using a decision-analytic model could identify what levels of extra-colonic follow-up would be optimal, and thus are worth considering, if extra-colonic follow-up does in fact offer a survival advantage in the manner hypothesized. That is, because of the low number of patients who might benefit, it is possible that aggressive extra-colonic surveillance would not be worth the costs even if it was beneficial. In this case, establishing evidence of efficacy would be a moot point.

Several reports have evaluated the cost-effectiveness of competing post-diagnosis endoscopic surveillance strategies aimed at preventing and detecting second primary CRCs.^{147 148} However, to date there have been few attempts to evaluate the cost-effectiveness of alternative extra-colonic-focused follow-up strategies for patients who undergo curative resection for locally-advanced (stage II and III) CRC. Using the results of one of the earliest meta-analyses, Renehan, et al., (2004)¹⁴⁹ compared ‘intensive surveillance’ with ‘minimal surveillance’ using a 5-year time-horizon (that of the trials). They estimated ICERs of about 3,500-5,500 British Pounds Stirling (2017). However, this is of limited value because it is based only on the earliest 5 trials in which recurrence rates were much larger. It also does not help settle the question of whether 1, 2, 3, or some other number of CT scans should be offered but instead groups together and compares a heterogeneous set of strategies. Finally, it is unlikely that cost estimates from the UK in 2002 are meaningful for US policymakers in 2018. Park, et al, (2001)¹⁵⁰ evaluated offering a CT + FDG PET scan in place of just a CT scan to patients with suspected recurrence due to elevated CEA levels. However, this addresses a different decision problem than that considered here. Rose, et al., (2014)¹⁵¹ described a simulation

model of CRC recurrence overlain with surveillance and expressed the intention to use it to identify optimal follow-up strategies. However, no such analysis has appeared in the literature to date. Lee-Ying, et al., (2017)¹⁵² evaluated intensive follow-up among patients with limited metastatic disease who undergo curative resection. However, these patients differ notably from stage II and III patients in that their risk of recurrence is much greater. Gazelle, et al., (2003)¹⁵³ evaluated the cost-effectiveness of surgical resection with a curative intent for diagnosed metachronous liver metastases but did not include the costs of routine surveillance for an entire cohort of patients following resection for primary CRC. The cost-effectiveness of post-diagnosis extra-colonic surveillance for locally-advanced CRC thus remains an open question.

In this paper, I use the microsimulation model described in Chapter 2 to conduct a cost-effectiveness analysis of alternative extra-colonic-focused follow-up strategies for locally-advanced CRC patients using a lifetime time horizon and taking the perspective of the US health care system. I consider six different strategies based largely on the most recent US guidelines. Again, the full guidelines are given in Table A.1. While there is some discrepancy among different organizations, they generally call for routine clinical visits and CEA testing every 3-6 months for 5 years as well as annual CT scans for 3-5 years for stage II and III patients.

METHODS

In this paper, I compare (i) no routine CEA or CT follow-up (no surveillance), (ii), a single CT scan at 12-18 months (as considered in the FACS trial), (iii) current US guideline-level CEA testing and a single CT scan at 12-18 months, and guideline-level CEA testing and annual CT scans for (iv) 2, (v) 3, and (vi) 5 years. The six different

strategies are depicted in Table 11. For all analyses in this paper, I assumed compliance would be 100%. I justify and comment on the importance of this assumption in the discussion section. I did not attempt to compare 4 and 5 years of annual CT to 3 and 4 years, respectively, because exploratory analyses suggested that the Monte Carlo sizes needed to acquire sufficiently precise estimates of the ICERs were prohibitively large.

For the main analyses, the different surveillance schedules were compared using the ICER in terms of costs per additional life-year. This is defined as the ratio of the incremental costs to the incremental effectiveness, comparing a more effective (and generally more expensive) intervention to a less effective (and generally less expensive) intervention. The analyses use a lifetime horizon and thus the incremental effectiveness is the increase in remaining life-expectancy (LE) in years. Moreover, the perspective of the US healthcare system is taken so the incremental cost is the increase in mean overall medical costs. The included costs are discussed in detail below. The ICER is properly interpreted as the cost we would pay for each additional year of LE achieved in virtue of upgrading from a standard intervention to a more effective yet generally also more expensive intervention.

COHORTS

As discussed in Chapter 2, stage II patients face better prognosis after clear-margin salvage surgery than do stage III patients, and this phenomenon was explicitly incorporated into the model. Moreover, as shown in Chapter 3, the model predicts that they would have a higher salvage rate than stage III patients (see Table 10). On the other hand, stage II patients also face a notably lower recurrence rate. Thus, *a priori*, it is unclear if and how optimal follow-up would differ between stage II and III patients.

Similarly, it is unclear if and how optimal follow-up might differ between colon and rectal cancer patients. As will be explained below, the two types of disease lead to different costs for salvage therapy and surveillance. In addition, stage III colon cancer patients might be expected to benefit more from intensive follow-up than stage III rectal cancer patients due to the greater recurrence rate among the former population. The same idea applies to stage II rectal cancer patients who face a greater recurrence rate than stage II colon cancer patients.

Finally, the age of the patients is relevant because the benefit conferred by surveillance is a small chance to notably postpone or even prevent death from cancer. In a population with higher background risk of mortality, this benefit will be less valuable. Accordingly, different surveillance schedules were thus evaluated separately by stage (II and III), location (rectal and colon) and age (60, 65, 70, and 75). Patients in the last age group face a greater risk of surgery-related mortality (5% vs 2%), but otherwise the age-specific cohorts differ with respect to the effectiveness components of the model only in their background mortality rates. In the youngest two cohorts (60 and 65 years old), 58% of patients were male, and in the other two cohorts 55% were male. This reflects the actual age-category-specific sex distribution among new cases of CRC in the US population.¹

For stage II and III rectal cancer patients (as described in Chapter 2) the simulated cohorts represent the disease experience of patients initially treated with adjuvant 5-FU/LV and possibly neoadjuvant radiation therapy and patients initially treated with both neoadjuvant chemoradiation therapy and adjuvant chemotherapy using a FOLFOX regimen, respectively. For the stage II and III colon cancer cohorts, however, it was

necessary to select the proportion of patients receiving adjuvant chemotherapy. The goal was to mimic the utilization that would occur among the cohort of US colon cancer patients of the corresponding stage and age who undergo routine follow-up. The assumed rates of utilization of each regimen are given in Table 12. The estimates for stage III patients were taken from tables provided by Professor Kuntz based on an analysis SEER-Medicare data (Personal Communication). While recent estimates in the literature suggest only about two-thirds of stage III colon cancer patients in the US actually receive adjuvant chemotherapy¹³⁶, these rates are higher among younger patients and those with less serious comorbidity¹⁵⁴ - the sort of patients who are more likely to undergo intensive surveillance. For stage II, estimates from the literature based on incident cases in the National Cancer Database from 1998-2006 put adjuvant utilization at around 40% for patients 65 and under, with about 2/3 of these patients receiving double-agent regimens.¹⁵⁵ Estimates of the utilization of any adjuvant chemotherapy among the 70- and 75-year-old stage II cohorts were taken from an analysis of incident cases in the SEER-Medicare linked database over the period of 1992-2005.¹⁵⁶ I assumed a smaller proportion of these patients used double-agent regimens (FOLFOX) since the previous reference¹⁵⁵ documented a notable decline among patients over 65.

Table 11: Follow-up Regimens under Investigation

Follow-up Arm	Test	3	6	9	12	15	18	21	24	30	36	42	48	54	60
Arm 1: No Follow-up	CT														
Arm 2: 1 CT	CT					X									
Arm 3: CEA + 1 CT	CT					X									

Arm 4: CEA + 2 CTs	CT				X				X						
Arm 5: CEA + 3 CTs	CT				X				X		X				
Arm 6: CEA + 5 CTs	CT				X				X		X		X		X
Arms 3-6	CEA		X	X	X	X	X	X	X	X	X	X	X	X	X

Table 11 depicts the schedules of CEA assays and CT scans for the 6 Arms of extra-colonic follow-up evaluated in this cost-effectiveness analysis. Arms 1 and 2 have no CEA follow-up.

Table 12: Utilization of Adjuvant and Palliative Chemotherapy by Cohort Age and Stage

Disease	Context	Age	FOLFOX	5-FU/LV (Fluoropyrimidine)	None
Stage III Colon	Adjuvant	60	90%	5%	5%
		65	75%	15%	10%
		70	60%	25%	15%
		75	35%	35%	30%
Stage II Colon	Adjuvant	60	27%	13%	60%
		65	27%	13%	60%
		70	18.5%	18.5%	63%
		75	9%	19%	72%
Unresectable Recurrence (Rectal and Colon)	Palliative	60	100%	0%	---
		65	100%	0%	---
		70	75%	25%	---
		75	50%	50%	---

Table 12 gives the percent of colon cancer patients assumed to use FOLFOX, 5-FU/LV, or no chemotherapy in an adjuvant setting by the age and stage. By adjuvant setting, I mean as part of the initial treatment for the primary CRC tumor. It also gives the percent of patients (colon or rectal) with unresectable recurrence treated with FOLFOX and a single-agent fluoropyrimidine regimen.

It was also necessary to make an assumption concerning the treatment (single-agent or double-agent cytotoxins) received by patients with unresectable recurrence (Table 12). I assume that all patients healthy enough to partake in routine surveillance who then develop unresectable recurrence are at least treated with a single-agent first-line regimen

of systemic chemotherapy, e.g., 5-FU, and that most are treated with a double-agent regimen, namely FOLFOX. Younger and healthier patients are more likely to be treated with double-agent chemotherapy^{157 158}, so I assumed all patients in the 60- and 65-year old cohort would be treated with FOLFOX. For the 70- and 75-year old cohorts, my assumption was based on estimates from an analysis of SEER-Medicare data using incident cases of stage IV CRC in 2009.¹⁵⁹

In order to increase the efficiency of the simulations, I used a variance-reduction technique. For each cohort e.g., 65-year-old stage III colon cancer patients, I simulated six nearly identical arms of patients who were exposed to different surveillance regimens. That is, I first defined a heterogeneous cohort of different patients who varied with respect to sex, heterogeneity¹⁶ terms, recurrence-status, time of death from other causes, etc. I then replicated this cohort five times to make a total of six treatment-arms. Thus every individual appeared in each of the six different treatment arms. Such replicated individuals were identical in terms of age at the start of the simulation, sex, adjuvant treatment received, heterogeneity terms, recurrence-status (yes or no), time to detectable and clinical disease (if they suffered a recurrence), time of death from other causes (based on lifetables), and whether they received 5-FU or FOLFOX in the event of unresectable recurrence. The only differences across replications of the same individual were the follow-up received and the results of any pseudorandom simulations that were event-dependent (with the exception of time to detectable and clinical disease). For example, if

¹⁶ As a reminder to the reader, there were 3 correlated heterogeneity terms at the individual level. These were for the time-to-detectable-disease process (latency time), the time-to-clinical-disease process (sojourn time), and the time-to-cancer-death processes.

multiple versions of the replicated individual presented with clinically-detected disease, it is possible one would have resectable disease while the others might not. Similarly, one might die of salvage surgery while another might be cured.

ANALYSES

For each age-, stage-, and location-specific cohort, the model simulates the (remaining) life course of patients in the six different treatment arms and calculates the cumulative costs of surveillance and treatment for recurrence (explained below). As noted in Chapter 2, time of death from other causes was simulated from the most recent sex-specific US lifetables available from the National Center for Health Statistics.¹⁶⁰ The probability of dying during the 100th year was readjusted so that no patients lived to be older than 100.

The effectiveness of each follow-up regimen is evaluated by the LE among all patients in the treatment-arm, and the incremental efficacy is equal to the increase in LE in moving to the more intensive intervention, i.e., adding another CT scan. Since the proportion of patients who benefit at the margins, i.e., with the addition of the last CT scan, is likely to be very small, the change in LE among the entire treatment arm is likely to be extremely small for some comparisons. This may lead readers to conclude that more intensive follow-up is not clinically relevant. However, this may not be the case. A moderate benefit that applies to only a (very) small proportion of patients will appear small when it is averaged over a large cohort. And certainly, because costs accrue to all patients who undergo the follow-up regimen, the *ex-ante* expected benefit of the intervention (averaged over the entire treatment arm) is the appropriate measure to compare with average costs in a decision analysis. However, a very small average benefit

should not be dismissed simply because of its size. Rather, that is the point of also calculating the mean incremental costs and calculating ICERs. In order to aid the reader in avoiding the previously mentioned natural yet problematic intuition, I also estimate the incremental mean benefit among patients who suffer a recurrence. That is, I calculate the increase in LE, from the beginning of the simulation, among patients who eventually suffer a recurrence. The point of calculating LE from the beginning of the simulation rather than from the time of recurrence is to avoid lead-time bias whereby patients with more intensive follow-up simply have their disease detected earlier. However, I only report the results of this calculation for 65-year old stage III colon cancer patients.

For the reference-case analysis, both life-years and costs were discounted at 3%. However, I also performed the analyses with a 0% discount rate. Because of the discrete-event simulation framework of the model, I used constant exponential discounting¹⁶¹ to adjust for the timing of costs and benefits. The instantaneous rate of discounting λ was backed out so that the value of the exponential discount function at exactly 1 year was equal to the discount factor at 1 year in a discrete-time discounting context using a discount rate of 3%, i.e., $\exp(-\lambda) = \frac{1}{(1+0.03)^1}$. In the case of life years (which accrue continuously), the discounted value of t additional life years (from the perspective of the present) is equal to the definite integral of the exponential discounting function from 0 to t , i.e., $\int_0^t \exp(-\lambda u) \partial u$. In the case of costs, the discounted value of a cost C occurring at a time t years in the future is equal to the product of the cost and the exponential discounting function evaluated time t , i.e., $C \exp(-\lambda t)$.

For each cohort (age, stage, and location), I ran the simulation multiple times and averaged over the results. Each iteration of the simulation required (pseudo)randomly drawing a set of parameters for the time-to-detectable-disease and time-to-clinical-disease processes for each included group of adjuvant treatment. As a reminder to the reader, the point of this was to try to average over the nonidentifiability of parameters governing the distribution of latency and sojourn times. In the case of rectal cancer, there was only one type of adjuvant treatment per cohort and thus only one distribution of parameters to sample from. But in the case of colon cancer, there were three types of adjuvant treatment per cohort. Again, the breakdown of these is given in Table 12. For convenience, for each iteration of the simulation, I sampled from the three parameter sets, e.g., stage III colon cancer patients treated with no adjuvant, 5-FU/LV, and FOLFOX, independently.

Monte Carlo sizes and the number of iterations used were selected based upon the magnitude of the effect size. To compare arms 2 and 3 to arms 1 and 2 (see Table 11), respectively, I simulated 25 iterations with a Monte Carlo size of 600,000 patients per arm in each iteration. To compare arms 3, 4, 5, and 6, I used 75 iterations per cohort and a Monte Carlo size of 1 million per arm per iteration. After the full simulation, the arm-specific mean costs and LEs were combined across iterations to form a set of overall arm-specific mean costs and LEs. The arms were then organized by increasing efficacy, and the appropriate incremental costs and incremental LEs were calculated. Finally, I calculated the corresponding ICERs.

In order to aid clinical decision-making, I used the ICER point estimates to identify the optimal follow-up regimen for each cohort under a range of values for a life-year. By

the value of a life-year I mean a policymaker's willingness to pay (WTP) for an additional expected life-year. In particular, I used increments of \$25,000 up to \$150,000 as the value of life-year and determined what follow-up regimen should be used for each cohort in each scenario. I did this both for the analyses using a discount rate of 3% and for the analyses using a discount rate of 0%.

By optimal I mean the most intensive follow-up regimen with an ICER less than or equal to the assumed value of a life-year. So for example, if the ICER comparing routine CEA and 2 CT scans to routine CEA and 1 CT scan was \$73,000 for a given cohort, a second scan would not be adopted with a WTP of \$50,000, but it would be adopted with a WTP of \$75,000 or more. However, it would only be considered the optimal follow-up regimen (among those under investigation) for a given WTP if there was not a more intensive follow-up regimen (e.g., CEA and 3 CT scans) with an ICER that was also less than or equal to the same WTP value.

The model described in Chapter 2 is a complex microsimulation model. Therefore, Monte Carlo sampling must be used to estimate moments or functions of moments for outcomes of interest. This differs from a cohort Markov model where analytic solutions for such moments exist.¹⁶² As with all Monte Carlo analyses, it is important to quantify the precision of estimates. I therefore calculated 95% confidence intervals (CI) for incremental LEs, incremental costs, and ICERs. This required estimation of empirical covariance matrices among arm-specific LEs and among arm-specific mean costs in each iteration. There was a very high degree of positive correlation among arms in each iteration because of the variance-reduction technique I used (simulating identical individuals). Covariance matrices for the overall arm-specific mean costs and LEs were

easily constructed from the iteration-specific covariance matrices since the iterations were independent. I used the delta method¹⁶³ to construct an estimate of the sample variance of each ICER. This in turn required estimation of the covariance between the LE and mean costs, both between and within the two compared arms. I note that, because of the variance-reduction technique I used, an estimate of the mean costs in one arm was correlated with an estimate of the LE of another arm (within a given iteration).

I note that the interpretation of confidence intervals for an ICER can be problematic when the upper and lower limits are not both positive. This is because a negative ICER can mean either that a more effective intervention is cost-saving or that a more costly intervention is less effective. However, when the entire 95% CI covers only a positive set, the interpretation is clear. It simply quantifies the precision of the ICER estimate comparing a more effective and costly intervention to a less effective and costly intervention.¹⁷ Accordingly, I only present ICER CIs when the entire interval is positive.

OVERVIEW OF COSTS

To conduct the cost-effectiveness analysis, costs were added to the microsimulation model described in Chapter 2. Because the analyses were undertaken from the perspective of the US healthcare system, they only include direct and indirect medical costs. Direct medical costs include costs related to follow-up testing while indirect medical costs include costs associated with treatment of recurrence. It is important to include the latter costs since the benefit of surveillance accrues specifically from this

¹⁷ Strictly speaking, a positive ICER could also result when an intervention is both less efficacious and less costly than the comparator (standard care). However, the interpretation is also clear in this case if the ICER 95% CI includes only positive values.

more intensive treatment of recurrence. I did not include medical costs unrelated to cancer nor medical costs related to treatment of the primary CRC. For example, the cost of the primary surgery and any adjuvant chemotherapy were not included. Nor were costs associated with post-treatment but pre-recurrence care for the primary disease.¹⁶⁴ The reason for this exclusion is convenience and the fact that these costs should be the same across arms. This is because intensive follow-up will only affect how recurrence is treated and potentially how long a patient will live after being diagnosed with recurrence. However, downstream costs associated with care for the patient related to treatment of recurrence were included. Finally, other patient-related costs (time and money spent traveling to care and time spent getting care), productivity costs, and most survivor costs were also not included since a societal perspective was not used. Though again, in the case of the survivor costs, CRC-related costs (those associated with the long-term treatment of recurrence) incurred in added years of life were included.

The two main categories of medical costs include surveillance-related costs and those related to treatment of recurrence. The resource utilization associated with surveillance was evaluated based on 2017 Medicare reimbursement rates as listed in the Center for Medicare & Medicaid Service's (CMS) Medicare Physician Fee Schedule (MPFS). Because routine follow-up testing generally occurs in an outpatient setting, I used the national average reimbursement rate for the combined technical and professional components of services. Services were identified by their Healthcare Common Procedure Coding System (HCPCS) codes – CMS's adaptation of the American Medical Association's Current Procedural Terminology (CPT) codes. The MPFS reimbursement rates represent the maximum allowable fee and include the amount that will be

reimbursed by Medicare as well as any amount which must be paid by the patient. I included the full cost amount and did not attempt to remove patient contributions, e.g., coinsurance or copays, since these would vary from patient to patient depending upon whether they have additional coverage, e.g., Medicaid.

In general resource use related to salvage treatment with a curative intent was also evaluated using Medicare reimbursement rates. Other than any neoadjuvant or adjuvant therapy, these resources would be consumed in an inpatient setting. They were thus evaluated using Medicare's inpatient prospective payment system for acute care hospital inpatient stays. This was done by determining an appropriate diagnosis-related group (DRG) for all such services. Each DRG includes a homogenous class of inpatient services and has been assigned a payment weight based on the estimated mean resource use required to treat Medicare patients included in that DRG. When combined with a national base payment rate, these weights determine the hospital reimbursement for non-physician-related labor, non-labor-related operating costs, and capital costs. For the national base payment rate, I used the value of \$5,962.93¹⁸. Physician costs associated with inpatient services were assigned based upon reimbursement rates listed in the MPFS for particular procedures identified by HCPCS codes. However, the reimbursement for these procedures does not include a technical component.

All remaining medical costs were taken from the literature. This was true for all costs related to treatment of patients with unresectable disease and for costs related to long-

¹⁸ This combines the 2017 federal capital rate and the 2017 labor and non-labor rates for a Hospital which submitted quality data and used Electronic Health Records (EHR's). According to CMS, over 50% of providers used HER's.

term care for patients surviving after salvage surgery. Cost values were inflated to 2017¹⁹ US dollars using the medical-component of the Consumer Price Index (CPI).¹⁶⁵

SURVEILLANCE-RELATED COSTS

The important surveillance-related costs are given in Table 13. These include the costs of an abdominal (abdominal/pelvic) and thoracic CT scan with IV contrast for colon (rectal) cancer patients and the cost of a CEA assay. I assume any such regimen would be offered over and above routine clinical visits to monitor the surgical wound and any other treatment-related morbidity and for the patient to report any symptoms. Thus, routine clinical visits were not included in the reference-case analysis since they are assumed to be the same across arms. However, I assumed that any positive test result would require an additional 30-minute clinical visit.

In addition, as described in Chapter 2, both a true-positive and false-positive test result incur additional costs over and above the additional 30-minute clinical visit. In the case of a CEA assay, a positive result leads to a CT scan (with reduced specificity). For a positive CT finding, I assumed the cost differs between a true positive and a false positive. In the case of a true positive CT scan, I assume 5% of patients undergo either a laparoscopy or thoracoscopy and 5% of patients get a CT-guided fine-needle aspiration (FNA). These would be needed in the case of equivocal CT findings before further treatment was provided. For rectal cancer patients, I further assume that 25% of true-positives undergo a pelvic MRI to confirm a pelvic finding on CT and/or to determine the

¹⁹ The medical-component of the CPI inflation rate was not available from the source I used for converting 2016 dollars into 2017 dollars. I therefore assumed 2017 would have the same rate as 2016.

appropriateness of pelvic salvage surgery. The expected additional cost of a true-positive CT finding (not including any additional 30-minute clinical visit) for rectal and colon cancer patients is given in Table 13.

In the case of a false-positive CT finding, I assume that patients would receive one of four further diagnostic procedures in equal proportion: (1) repeat CT scan (of the abdomen or chest), (2) an MRI (of the pelvis or abdomen), (3) a CT-guided FNA, or (4) a laparoscopy or thoracoscopy. The corresponding expected costs of a false-positive CT finding for rectal and colon cancer patients are also given in Table 13. The relevant HCPCS codes and reimbursement rates for individual procedures are given in Table D.I. The construction of the aggregated costs is illustrated in Table D.3.

Table 13: Costs of Surveillance and Treatment of Recurrence

SURVEILLANCE COSTS		
Clinical Visit (30 Min)		\$109
CEA Assay		\$26
Full CT Scan for Rectal Patients		\$548
Full CT Scan for Colon Patients		\$467
Cost of a True Positive CT: Colon		\$40
Cost of a True Positive CT: Rectal		\$144
Cost of a False Positive CT: Colon		\$374
Cost of a False Positive CT: Rectal		\$369
INITIAL COST OF SALVAGE TREATMENT		
Cost of Failed/R2 Resection: (Colon)		\$12,215
Cost of Failed/R2 Resection: (Rectal)		\$11,287
Expected Cost of R0-1 Surgical Resection + (Neo)Adjuvant Therapy (Colon)		\$54,442
Expected Cost of R0-1 Surgical Resection + (Neo)Adjuvant Therapy (Rectal)		\$45,192
DOWNSTREAM COSTS OF SALVAGE TREATMENT		
Expected Cost of Follow-up CT Imaging Per Round of Follow-up (Colon)		\$486
Expected Cost of Follow-up CT Imaging Per Round of Follow-up (Rectal)		\$567
Expected Cost of Possible Second Surgical Resection with Curative Intent + (Neo)Adjuvant Therapy (Colon)		\$9,255
Expected Cost of Possible Second Surgical Resection with Curative Intent + (Neo)Adjuvant Therapy (Rectal)		\$6,779
Age-Specific Monthly Costs for Patient Care	< 75	\$1,093
	75-85	\$656
	≥ 85	\$473
Monthly Costs for Terminal Cancer		\$4,091

COSTS OF PALLIATIVE TREATMENT		
Monthly Costs for Initial Treatment		\$7,438
Age-Specific Monthly Costs for Patient Care	< 75	\$3,683
	75-85	\$2,210
	≥ 85	\$1,473
Monthly Costs for Terminal Cancer		\$4,091

Table 13 gives the aggregate costs used in the cost-effectiveness analyses for surveillance tests, salvage treatment, and treatment for patients with unresectable recurrence ('palliative treatment'). Values are in 2017 US dollars. Future costs have not been discounted.

COSTS OF TREATMENT WITH A CURATIVE INTENT

Resource use related to treatment with a curative intent can be divided into two categories: the initial treatment and long-term care. The initial treatment involves salvage surgery, treatment for any resulting serious morbidity (e.g., increased inpatient stay), and possibly perioperative radiation or chemotherapy. Long-term care involves additional intensive surveillance, the possibility of a second salvage surgery, routine care, and treatment of terminal disease among those who ultimately suffer a re-recurrence and die from the disease.

I evaluated resource use for initial treatment separately depending on the margin status of surgery. Patients who undergo a failed or R2 resection are only assigned costs associated with the procedure. They are then considered to have unresectable disease and are assigned costs accordingly (discussed in the next section). By failed resection I mean a laparotomy or thoracotomy that does not proceed to surgical resection due to the discovery of previously undetected widely-disseminated disease. An R2 resection refers to a macroscopically incomplete resection of the tumor. I combine the two in the model since they are assumed to lead to similar outcomes, i.e., no benefit. For the cost analysis, these patients are assumed not to be susceptible to serious morbidity or mortality in the way that patients who undergo R0-R1 resections are. These latter patients undergo a more

significant procedure and generally lose a potentially significant portion of vital organs. This is generally not true of the former patients.

For patients who undergo an R0-R1 resection (macroscopically clear margin), I evaluated resource use separately by anatomical location of salvage surgery and then used a weighted average of these costs to represent the expected cost of salvage treatment with curative intent. That is, for each anatomical location of recurrence (e.g., liver, lung), I calculated an expected cost of salvage treatment which included the costs of surgery, extra costs due to mortality and morbidity, and the costs of radiation therapy and chemotherapy. I then combined these into an overall expected cost of (macroscopically-clear margin) salvage treatment of recurrence for both rectal and colon cancer patients. For colon cancer patients, I distinguished between hepatic resection, pulmonary resection, and abdominal resection of local-regional recurrence. For rectal cancer patients, I distinguished between hepatic resection, pulmonary resection, and pelvic resection. Thus, there were 4 distinct types of salvage treatment to consider.

In the case of resection of local-regional recurrence in the abdomen for colon cancer patients, I used an estimate of the mean cost of surgery for stage III colon cancer patients taken from the linked SEER-Medicare database as a proxy.¹⁶⁶ Salvage surgery for abdominal recurrence among colon cancer patients is generally extensive and often involves multivisceral resection.¹⁰⁶ I selected surgery for stage III patients since it was the most expensive among non-metastatic patients. The estimate was based on a 3-month period starting in the month of diagnosis and so is unlikely to include adjuvant costs. Moreover, preoperative and intraoperative radiation therapy and neoadjuvant chemotherapy are not common practice in the case of surgical resection of primary colon

cancer tumors, but they are used in the case of salvage therapy (for local-regional recurrence) due to the difficulty of achieving a complete resection. I thus added costs associated with intraoperative radiation therapy and neoadjuvant and adjuvant radiation therapy and chemotherapy. The relevant assumptions and aggregate costs are given in Tables D.2 and D.3, respectively.

For hepatic, pulmonary, and pelvic resections, I identified appropriate DRGs and multiple HCPCS codes. In the case of pelvic resection, separate DRGs were assumed for male and female patients and a weighted average was used. For hepatic and pulmonary resections, there were three potentially-relevant DRGs depending on if the procedure involved major complications or comorbidities (MCC), complications or comorbidities (CC), or neither (None). In the case of pelvic salvage, there were only two relevant DRGs (CC vs. None). A previous cost-effectiveness analysis of the resection of metachronous hepatic metastases in CRC patients estimated that 12% of cases involved no complications or comorbidities.¹⁵³ In the absence of more recent information, I used this assumption for all 3 types of salvage surgery, but it was estimated prior to Medicare's adoption of two categories for complication and comorbidities (MCC vs CC). Unfortunately, a literature search was not helpful on the distribution of cases with respect to these categories, so I assumed that 50% of patients with complication or comorbidities would be assigned as MCC and 50% as CC (for hepatic and pulmonary resection).

For physician reimbursement, I identified several different HCPCS codes for each type of salvage surgery. These were selected and assigned weights to reflect the distribution of procedures reported in relevant case series. The relevant assumptions and sources are given in Table D.2, and the codes and reimbursement rates are given in Table

D.1. Aggregate costs are given in Table D.3. In the case of failed or R2 resections, I followed the previously mentioned analysis¹⁵³ and assumed patients were assigned the lowest-cost level DRG (no CC) and used a Laparotomy or Thoracotomy for physician reimbursement.

Estimates of the costs associated with surgical morbidity and mortality were taken from the literature. These costs represented costs that accrued over and above the costs of the initial surgical procedure. They were incorporated into the expected cost of salvage treatment²⁰ by multiplying the estimated mean cost per case times the probability of serious morbidity and mortality. In the case of hepatic resection, estimates were available for the mean cost per case of serious morbidity and per case of 30-day mortality. The latter was more than twice the cost of the former. In the case of pulmonary and pelvic salvages, only estimates of the average cost per case of serious morbidity were available. In the absence of better information, I therefore used 200% of the average cost per case of serious morbidity as an estimate for the average cost per case of treatment-related mortality.

A detailed breakdown of procedures and reimbursement rates or costs from the literature for each type of salvage treatment is given in Table D.1. This includes the mean added cost per case of serious morbidity and mortality. Table D.3 gives the aggregated costs and illustrates the appropriate calculations. Table D.2 summarizes important auxiliary assumptions and gives the relevant sources in the literature.

²⁰ This was not done in the case of salvage therapy for local-regional recurrence in the abdomen since the average cost taken from the literature for surgery for stage III colon cancer already included such costs.

The last component of the initial treatment costs was preoperative and postoperative external-beam radiation therapy and chemotherapy. Estimates of the rates of utilization and costs of perioperative radiation therapy and chemotherapy were taken from the literature. The incremental effectiveness of these treatments over and above salvage surgery remains unclear¹⁶⁷, but they are routinely used in a subset of patients and represent a major cost. An estimate of the mean cost of neoadjuvant radiation therapy was taken from an analysis using SEER-Medicare data.¹⁶⁸ For chemotherapy, the main distinction in the literature is between a double-agent and single-agent form of chemotherapy. I assumed all patients using double-agent (single-agent) chemotherapy would use FOLFOX (5-FU/LV). For each type of salvage treatment, assumptions about the proportion of patients using chemotherapy, the proportion treated preoperatively and/or postoperatively, the type of chemotherapy used, and the assumed length of time patients undergo treatment are given in Table D.2. The estimated costs per 3 months of treatment (and sources) are listed in Table D.1. Finally, Table D.3 lists the *ex-ante* expected cost of (neo)adjuvant (chemo)radiation therapy per patient for each anatomical location of salvage treatment.

Finally, a weighted average was calculated of the expected costs for the different anatomical locations of salvage treatments. These values are given in Table 13. Unfortunately, the intensive-follow-up trials did not generally provide detailed information about the distribution of types of salvage surgeries, and these likely have changed since the earliest trials. I therefore estimated weights from a large cancer-registry-based study of the distribution of isolated recurrences among colon and rectal cancer patients.⁴ For colon cancer, I assumed 60% and 20% of cases were liver and lung

resections, respectively, and 5% of patients involved staged resection of both liver and lung metastases. This is apparently a viable form of salvage treatment.¹⁶⁹ The remaining 15% of patients were assumed to undergo salvage for local-regional recurrence located in the abdomen. For rectal cancer, I assumed 40% and 30% were liver and lung resections and an additional 5% underwent staged resection of both hepatic and pulmonary metastases. The remaining 25% were assumed to undergo salvage treatment for pelvic recurrence.

Two of the costs characterized above as part of long-term care relate to detection and treatment of re-recurrence. The microsimulation model described in Chapter 2 does not currently model the re-recurrence process or a re-salvage process. It therefore does not model surveillance and detection of re-recurrence. However, it is important that these costs are included in the analyses. Thus, all patients who undergo a salvage surgery are assumed to partake in intensive routine CT imaging – every 6 months for 2 years and then annually for 3 years – while alive and with a cancer-specific life-expectancy of at least 10 months. Routine CEA measurement, however, is not included in this schedule. The 10-months figure was selected because, in the model, the last 9 months are considered a terminal phase during which patients are treated the way patients with unresectable disease are treated at the end of their life (discussed below). It is also assumed that 15% and 17% of rectal and colon cancer patients, respectively, undergo a second attempt at salvage surgery.^{11 170-172} The expected costs per CT scan completed (including false-positives) and of a second salvage surgery are given in Table 13.

Another major component of long-term costs is referred to as ‘continuing costs’ in the literature. Patients who undergo surgical resection for CRC (of any stage) continue to use

more medical resources than comparable cancer-free patients for at least 10 years.¹⁶⁴ That is, between the periods of initial treatment and any terminal stage during which the patient dies of cancer, additional medical resources are required to care for the patient because of their CRC diagnosis. Continuing costs are greater for younger patients and higher stage patients per unit of time.¹⁶⁴ Most of the patients in the microsimulation model who undergo salvage treatment with a curative intent would have isolated metastatic disease. Continuing costs for such patients are likely less than for patients with unresectable disease. For the reference case, I therefore used an estimate of the continuing costs for stage III CRC patients as a proxy for the costs of continuing care after initial treatment with salvage surgery. These costs are listed in Table 13. The reference for these estimates used SEER-Medicare data for patients diagnosed with CRC between 1996-2002. It is unlikely that many patients were undergoing routine CT surveillance during this period and thus that such costs would be included in the estimate. Therefore these continuing costs are assumed to apply over and above the costs of a second round of intensive follow-up discussed above. I also assume that costs of possible salvage therapy for recurrence are not included.

In the microsimulation model, continuing costs accrue beginning in the 7th month after the salvage surgery until the patient's cancer-specific life-expectancy drops below 10 months or until 10 years have passed from the date of salvage surgery, whichever occurs first. I assume that the first 6 months represent the initial treatment period during which the costs of salvage therapy discussed above are borne. Moreover, the last 9 months of a patient's life (among patients who die from cancer) are treated separately.

Finally, all patients who die of cancer (i.e., suffer a re-recurrence and die from it) are assumed to have the same experience at their end of life as patients who are initially treated for unresectable recurrence. During this final phase of life (the ‘terminal’ phase), patients accrue costs at a faster rate (Table 13).^{164 173} The costs presented in Table 13 are assumed to include any palliative procedures as well as any chemotherapy given in the last stage of life. The terminal phase lasts for up to 9 months and takes priority over continuing care in the assignment of months. That is, if a patient lives 15 or fewer months after salvage surgery, only terminal phase costs are assigned; the patient never accrues continuing costs. In addition, if the patient lives fewer than 15 months after salvage, terminal months are only assigned after 6 months (of initial treatment). It is assumed the patient simply has a very aggressive re-recurrence.

COSTS OF TREATMENT FOR UNRESECTABLE DISEASE

I assume patients in the microsimulation model who present with unresectable recurrence incur costs at a rate comparable to patients with stage IV disease. In the literature, these costs are generally divided into three phases of the disease: initial, continuing, and terminal.^{164 173} Estimates of the monthly costs for each phase (and their source) are given in Table 13. The initial treatment phase is assumed to last up to 8 months and to include at least the first line of chemotherapeutic treatment. The costs of the terminal phase are identical to those used for patients treated with curative resection but who ultimately die from cancer. They accrue for up to the last 9 months of the patient’s life. In the case of unresectable recurrence, terminal months are prioritized over initial treatment months in that up to the first 9 months are classified as terminal disease before any initial treatment month are assigned. Finally, any additional months of lifetime

are assigned costs at the relevant age-specific rate for continuing costs (for stage IV patients). This was the strategy used to estimate the monthly costs in the selected source. Moreover, while perhaps unconventional, the durations of 8 and 9 months for the initial and terminal phases, respectively, were selected because they were the reported average amount of time spent by contributing patients in those phases. This is relevant because the authors reported average aggregate costs for the initial and terminal phases rather than monthly costs.

It should be noted that these values were estimated with SEER-Medicare data for patients diagnosed with stage IV CRC between 1996-2002.¹⁶⁴ I was unable to find a more recent US population-level source which estimated the excess costs of CRC for stage IV patients. It is very likely that these costs will have increased since the index period, potentially significantly, due to the increasing complexity of treatment for unresectable disease (e.g., the adoption of targeted therapy such as biological agents).¹⁷⁴ However, in the event that this does bias results, it is likely to be in favor of less intensive follow-up. All patients who die of cancer (from recurrence) will incur these costs, regardless of whether they undergo salvage surgery. However, patients who are cured from salvage surgery are spared these costs. This is one way in which more intensive upfront resource uses – the costs of salvage treatment – may lead to a decrease in downstream CRC-related medical costs. Thus, a larger value associated with the cost of terminal disease would mean greater cost-savings for successful salvage therapy.

SENSITIVITY ANALYSES

To assess the robustness of the simulation results to parameter uncertainty, I performed a set of univariate sensitivity analyses. These evaluated the effect of variation

of a single cost value. I focused on costs which were deemed both potentially particularly important and/or particularly uncertain. These included the initial costs of salvage treatment, the cost of palliative care, the cost of routine care (continuing costs) after salvage treatment, the cost of a CT scan, the cost of a true positive and false-positive CT result, and the cost and utilization of clinical visits. The point of this set of analyses was to evaluate the robustness of the results of the decision analysis (i.e., identification of optimal follow-up for a given WTP) to uncertainty surrounding these particular parameters considered in isolation (that is ignoring uncertainty among other parameters) and to assess how important these parameters are in determining the results. They were not meant to represent an exhaustive sensitivity analysis. I therefore only performed these analyses using the cohort of 65-year old stage III colon cancer patients. This cohort could reasonably be considered the most representative cohort. For each analysis, I simulated 15 iterations of the model with a Monte Carlo size of 600,000 per arm (and thus 3.6 million per iteration). Results were combined across iterations as they were for the reference-case analyses described above.

For the first two (salvage and palliative treatment), I ran the analyses using values that were 50% and 150% of the best-guess parameters described above. For example, instead of \$54,442 as the mean initial cost of salvage treatment for colon cancer patients, I used values of \$27,221 and \$81,663, respectively. Given the number of assumptions that went into calculating the average cost of salvage treatment, it was unrealistic to assess the impact of all of the assumptions or even just several of them. Rather, I chose to assess the impact of variation in the aggregate estimate. I also adjusted the expected cost of a second round of salvage treatment accordingly. In the case of the palliative care, I

adjusted all five costs associated with palliative care (bottom of Table 13) by 0.5 or 1.5. Because the cost of treating dying patients in the last 9 months of their life was assumed to be the same among patients who suffer an incurable re-recurrence after salvage treatment as it is among patients who initially present with unresectable recurrence, I also adjusted the monthly terminal costs for the former patients (those treated with attempted cure who die of cancer).

In the case of the continuing costs associated with salvage treatment, I ran an analysis using the continuing costs associated with stage IV patients (as opposed to stage III patients), i.e., the same values used for patients who present with unresectable recurrence and survive long enough to incur such costs. As can be seen in Table 13, the continuing costs reported in the literature for stage IV patients are notably larger than those for stage III patients. I also ran an analysis using continuing costs equal to only 50% of those estimated for stage III patients. This would be relevant if (a), as assumed for the reference-case analysis, an estimate of the continuing costs for stage III patients was more appropriate for patients after salvage therapy than was an estimate based on stage IV patients (the vast majority of whom would have had unresectable disease) and (b) the estimate from the literature of continuing costs for stage III patients was biased up (with respect to the model's purposes) because, contrary to my assumption, it included costs associated with routine surveillance and/or salvage therapy for recurrence.

I also performed a sensitivity analysis on three surveillance-related cost parameters. While Medicare reimbursement rates for an outpatient CT scan are available in the MPFS, it is potentially informative to evaluate how sensitive the results are to this cost. In particular, this may be relevant for patients with private health insurance where the

reimbursement for a CT scan is likely greater. I therefore ran the simulation using values 0.5 and 2.0 times the cost listed in Table 13 for colon cancer patients (\$234 and \$934, respectively). I also re-ran the simulation using notably larger values for the cost of a true-positive and false-positive CT scan. This would result from a higher proportion of patients requiring further diagnostic procedures and/or services before a treatment plan was made. For the average cost of a true positive, I used a value (\$400) 10x that listed in Table 13. I remind the reader that the original value (\$40) was based on several assumptions, including that only 10% of patients would require further diagnostic work. For a false positive, I used 2x the cost given in Table 13. Finally, I ran the analyses including the cost of a 60-minute clinical visit (\$209) for every surveillance test (using only 1 if both CEA and CT were scheduled) and after every positive test result. This differed from the reference-case analysis which only included the cost of a 30-minute clinical visit (\$109) in the case of a positive test result, assuming that patients would still attend routine clinical visits in the absence of routine extra-colonic focused surveillance.

I also evaluated the importance of the salvage-surgery-related mortality rate and the assumed regimen of systemic chemotherapeutic treatment for patients with unresectable recurrence – 5-FU vs FOLFOX – to the results. For the former, I used a 30-day mortality rate of 5%. In the case of the latter, I did this by re-simulating the same cohort (65-year old stage III colon) under the assumption that all such patients were treated with 5-FU (as opposed to the reference case where they all receive FOLFOX). This was deemed relevant because, as noted in Chapter 2, the calibrated median survival among patients with symptomatic unresectable recurrence who are treated with FOLFOX was 15.8 months. This figure drops to 10.4 months when patients are treated with 5-FU only. Since

this represents the expected survival experience of patients treated with salvage surgery in the counterfactual reality in which they were not treated with salvage surgery, a difference of this magnitude could potentially have a large influence on the results.

RESULTS

EFFECTIVENESS

Among patients who underwent no routine surveillance, undiscounted LEs from the start of the simulation varied from about 8 to 20 years depending on the starting age and the stage and location of the primary disease. Relative to baseline LE, changes in LE associated with more intensive surveillance regimens (incremental effectiveness) were small. The full effectiveness results are given in Table D.4. Below I briefly describe the results for the undiscounted analyses as they are generally more intuitively meaningful. While I report incremental effectiveness estimates in the unit of days, the ICER calculations used the unit of life-years.

With one or two exceptions, LE increased with each incremental component of surveillance. Adding a single CT scan added about 15-50 days of LE. Adding routine CEA follow-up in accordance with US guidelines further increased LE by about 27-103 days. The incremental benefits from further increases in surveillance intensity were very small. Adding a second (third) CT scan increased LE by only 2-11 days (0-8 days). Finally, adding both a 4th and 5th CT scan increased LE by 0-10 days. Among 75-year-old stage III rectal and colon cancer patients, adding both a 4th and 5th CT scan did not increase LE. The same was true among 75-year-old stage III colon cancer patients when adding a 3rd CT scan. These were the only cohorts and comparisons with a point estimate for the incremental effectiveness of 0 or less.

Overall, a few further patterns were observed among both the discounted and undiscounted effectiveness results. First, for each stage and location combination, e.g., stage III colon, the incremental benefit of any given upgrade, e.g., from 1 to 2 CT scans, decreased in magnitude as the age of the cohort increased. That is, controlling for disease type, a smaller benefit accrued with increasing age. Second, among the 60-, 65-, and 70-year-old cohorts, stage II rectal cancer patients almost always exhibited the greatest incremental benefit from adopting more intensive surveillance. In general, stage III and stage II colon cancer patients tended to have the 2nd and 3rd largest benefits on average, respectively. However, this result was not replicated among the oldest cohort.

When averaged over only those patients who suffered a recurrence, as opposed to the entire treatment arm, the undiscounted incremental benefits (gains in LE) associated with intensifying surveillance for 65-year-old stage III colon cancer patients were about 2 to 3.5 times as large. For example, patients who undergo a CT scan at 12-18 months and who suffer a recurrence could expect to live roughly 3.6 months longer than if they had not undergone the CT scan. Averaged across all patients in the treatment arm, this same figure is only about 1.3 months. Similarly, by adhering to a schedule of routine CEA testing in addition to the 1 CT scan, those who will suffer a recurrence can expect to live an additional 6.4 months compared to just undergoing the 1 CT scan. The analogous figure averaged across the entire treatment arm is 2.3 months. Although the expected incremental benefits of further enhancements to the surveillance schedule are larger when averaged only among patients who ultimately get a recurrence, they are still very small. By adhering to routine CEA testing and annual CT scans for 2 (3) years, (65-year-old stage III colon cancer) patients who will suffer a recurrence can expect to live an

additional 2.5 weeks (1 week) compared to if they had just undergone routine CEA testing and 1 (2) CT scans. Finally, patients who suffer a recurrence and adhere to 5 years of annual CT (and routine CEA testing) can expect to live almost another week compared to if they had just undergone 3 years of annual CT (and routine CEA). I remind the reader that these results are meant only to be informative and are not used in the cost-effectiveness analyses.

COSTS

The undiscounted (discounted) mean overall costs for patients who underwent no routine surveillance ranged from about \$12,000-\$40,000 (\$11,000-\$37,000), depending on the cohort. For these patients, the costs represent resource use associated with treating unresectable recurrence and, for a very small percentage of patients who suffered a recurrence, with salvage treatment. I note again that only resource-use related to recurrence was included in these analyses (since the patients did not undergo extra-colonic surveillance of any sort) and that total costs were averaged over the entire group of patients, including the up to 85 percent of patients who might not suffer a recurrence. Thus, while appropriate for a cost-effectiveness analysis, these values are not representative of a typical patient's recurrence-related costs.

The full cost results are presented in Table D.5. Relative to the baseline mean costs, the undiscounted (discounted) incremental costs associated with intensifying surveillance were quite a bit smaller, but definitely not trivial. The first CT scan added about \$2,100-\$4,900 (\$2,000-\$4,700) per person on average. Adding routine CEA testing added another \$3,900-\$6,700 (\$3,700-\$6,400) per person. This particular increase was substantial and generally constituted roughly 15-25% of the baseline costs among those

with no surveillance. Further additions lead to notably smaller incremental costs. A second CT scan added another \$600-\$900 (\$600-\$900) per person, and a third scan added between another \$500-\$1,000 (\$500-\$900) per person on average. Finally, the expected incremental costs associated with the 4th and 5th CT scans ranged from \$800 to \$1,800 (\$700 to \$1,600).

In general, increases in the costs associated with treating recurrence were the biggest drivers of the incremental costs associated with more intensive surveillance. That is, the increase in the rate of salvage treatment associated with more intensive surveillance lead to an increase in the average cost of treating recurrence (both in the short term and long term), and this increase was notably larger in magnitude than the level of resources consumed for actual surveillance. This was particularly true with the addition of the first CT scan and with the adoption of routine CEA testing. For colon (rectal) cancer, only about 13-15% (18-20%) of the incremental costs for these two comparisons was due to added surveillance tests. The rest was due to the increased utilization of salvage therapy. However, as surveillance regimens became more intensive, the difference in magnitude between these two components of the incremental costs diminished. In adding both a 4th and 5th CT scan, the incremental costs associated with surveillance were greater than the incremental treatment costs. This pattern was observed among all age-groups. In some cohorts, this phenomenon begun with the addition of the 3rd CT scan.

As with the effectiveness results, there were some discernable patterns in the cost results. With the exception of the 75-year-old cohort, stage III colon cancer patients had the greatest incremental mean costs with the addition of the first CT scan, routine CEA testing, and a second CT scan. However, this was not generally the case with the addition

of the 3rd CT scan and particularly not with the addition of the 4th/5th CT scan. This is likely because stage III colon cancer patients tend to have the shortest time to recurrence. In addition, with the exception of the addition of the first CT scan at 12-18 months, stage II rectal patients tended to have greater incremental costs than stage II colon cancer patients and stage III rectal patients among the younger 3 cohorts. In the oldest cohort, stage III rectal cancer patients generally had the highest incremental mean costs. Moreover, there was no clear pattern of decreasing (or increasing) incremental costs with increasing age as there was for the incremental benefit. That is, controlling for stage and disease location, older age did not necessarily mean greater or smaller incremental costs the way it meant smaller increases in LE.

COST-EFFECTIVENESS

The full set of ICER results are given in Table D.6. It is worth noting a few trends. First, discounted ICERs were categorically greater than undiscounted ICERs. Second, within a given cohort and comparison, e.g., routine CEA and 2 CT scans vs. routine CEA and 1 CT scan among stage III colon cancer patients, ICERs generally increased with age. That is, the cost per additional life-year achieved by upgrading to a more intensive surveillance schedule increased with age. Third, the two 75-year-old stage III cohorts were the only cohorts in which strategies were dominated. In particular, in the reference-case analysis, 1 CT scan at 12-18 months was dominated by extension by the combination of routine CEA testing and 1 CT scan for stage III rectal patients. Moreover, annual CT scans for 5 years was dominated by annual CT scans for 3 years for both stage III rectal and colon cancer patients as the former did not confer any survival advantage.

Estimates of the ICERs comparing routine CEA testing and 3 (5) annual CT scans vs. just 2 (3) annual CT scans were often very imprecise, particularly among older patients. This is because at a discount rate of 3% the already very small estimated incremental gains in LE from adding a 3rd or both a 4th and 5th CT scan among older patients flirted with negligibility. This in turn is due to the extremely small proportion of patients affected by such a change in follow-up care. However, in general the categorization of optimal follow-up was robust to this imprecision. Given this fact, it is unclear that much would be gained by undertaking the significant computational effort that would be required to achieve precise estimates of these ICERS.

Table 14: Optimal Follow-up by Cohort and Value of a Life-Year

COHORT	VALUE OF A LIFE-YEAR					
	\$25,000	\$50,000	\$75,000	\$100,000	\$125,000	\$150,000
Discount Rate = 0%						
Stage III Colon 60	1 CT	CEA + 2 CTs	CEA + 3 CTs	CEA + 5 CTs	CEA + 5 CTs	CEA + 5 CTs
Stage III Rectal 60	CEA + 1 CT	CEA + 2 CTs	CEA + 2 CTs	CEA + 3 CTs	CEA + 5 CTs	CEA + 5 CTs
Stage II Colon 60	CEA + 1 CT	CEA + 2 CTs	CEA + 3 CTs	CEA + 3 CTs	CEA + 5 CTs	CEA + 5 CTs
Stage II Rectal 60	CEA + 1 CT	CEA + 3 CTs	CEA + 5 CTs	CEA + 5 CTs	CEA + 5 CTs	CEA + 5 CTs
Stage III Colon 65	None	CEA + 2 CTs	CEA + 2 CTs	CEA + 3 CTs	CEA + 3 CTs	CEA + 5 CTs
Stage III Rectal 65	None	CEA + 2 CTs	CEA + 2 CTs	CEA + 2 CTs	CEA + 3 CTs	CEA + 3 CTs
Stage II Colon 65	1 CT	CEA + 2 CTs	CEA + 2 CTs	CEA + 3 CTs	CEA + 3 CTs	CEA + 3 CTs
Stage II Rectal 65	CEA + 1 CT	CEA + 3 CTs	CEA + 3 CTs	CEA + 5 CTs	CEA + 5 CTs	CEA + 5 CTs
Stage III Colon 70	None	CEA + 1 CT	CEA + 2 CTs	CEA + 2 CTs	CEA + 2 CTs	CEA + 3 CTs
Stage III Rectal 70	None	CEA + 1 CT	CEA + 2 CTs	CEA + 2 CTs	CEA + 2 CTs	CEA + 2 CTs
Stage II Colon 70	None	CEA + 1 CT	CEA + 2 CTs	CEA + 2 CTs	CEA + 3 CTs	CEA + 3 CTs
Stage II Rectal 70	None	CEA + 2 CTs	CEA + 2 CTs	CEA + 3 CTs	CEA + 5 CTs	CEA + 5 CTs
Stage III Colon 75	None	CEA + 1 CT	CEA + 2 CTs	CEA + 2 CTs	CEA + 2 CTs	CEA + 2 CTs

Stage III Rectal 75	None	CEA + 1 CT	CEA + 1 CT	CEA + 2 CTs	CEA + 2 CTs	CEA + 2 CTs
Stage II Colon 75	None	CEA + 1 CT	CEA + 1 CT	CEA + 2 CTs	CEA + 2 CTs	CEA + 2 CTs
Stage II Rectal 75	None	CEA + 1 CT	CEA + 1 CT	CEA + 2 CTs	CEA + 3 CTs	CEA + 3 CTs
Discount Rate = 3%						
Stage III Colon 60	None	CEA + 2 CTs	CEA + 2 CTs	CEA + 3 CTs	CEA + 3 CTs	CEA + 3 CTs
Stage III Rectal 60	None	CEA + 1 CT	CEA + 2 CTs	CEA + 2 CTs	CEA + 2 CTs	CEA + 2 CTs
Stage II Colon 60	None	CEA + 2 CTs	CEA + 2 CTs	CEA + 3 CTs	CEA + 3 CTs	CEA + 3 CTs
Stage II Rectal 60	1 CT	CEA + 2 CTs	CEA + 3 CTs	CEA + 5 CTs	CEA + 5 CTs	CEA + 5 CTs
Stage III Colon 65	None	CEA + 1 CT	CEA + 2 CTs	CEA + 2 CTs	CEA + 2 CTs	CEA + 2 CTs
Stage III Rectal 65	None	CEA + 1 CT	CEA + 2 CT	CEA + 2 CT	CEA + 2 CTs	CEA + 2 CTs
Stage II Colon 65	None	CEA + 1 CT	CEA + 2 CTs	CEA + 2 CTs	CEA + 2 CTs	CEA + 3 CTs
Stage II Rectal 65	None	CEA + 1 CT	CEA + 2 CTs	CEA + 3 CTs	CEA + 3 CTs	CEA + 5 CTs
Stage III Colon 70	None	None	CEA + 1 CT	CEA + 2 CTs	CEA + 2 CTs	CEA + 2 CTs
Stage III Rectal 70	None	1 CT	CEA + 1 CT	CEA + 2 CTs	CEA + 2 CTs	CEA + 2 CTs
Stage II Colon 70	None	CEA + 1 CT	CEA + 1 CT	CEA + 2 CTs	CEA + 2 CTs	CEA + 2 CTs
Stage II Rectal 70	None	CEA + 1 CT	CEA + 2 CTs	CEA + 2 CTs	CEA + 3 CTs	CEA + 3 CTs
Stage III Colon 75	None	None	CEA + 1 CT	CEA + 1 CT	CEA + 2 CTs	CEA + 2 CTs
Stage III Rectal 75	None	None	CEA + 1 CT	CEA + 1 CT	CEA + 1 CT	CEA + 2 CTs
Stage II Colon 75	None	CEA + 1 CT	CEA + 1 CT	CEA + 2 CTs	CEA + 2 CTs	CEA + 2 CTs
Stage II Rectal 75	None	CEA + 1 CT	CEA + 1 CT	CEA + 1 CT	CEA + 2 CTs	CEA + 2 CTs

Table 14 identifies the optimal surveillance schedule by cohort, discount rate, and the value of a life-year. Each cell lists the most intensive regimen of surveillance (for the corresponding cohort) that would provide additional life-years at a cost less than or equal to the assumed value of a life-year. In the event that the ICER for even a single CT scan is greater than the assumed value of a life-year, the optimal regimen is no surveillance, i.e., None. The top (bottom) section contains results using a discount rate of 0% (3%).

Table 14 gives the optimal surveillance schedule for each cohort of patients under a range of WTP thresholds, using both 0% and 3% discount rates. For the reference-case

analysis (discount rate = 3%), no surveillance would be optimal in all cases (except stage II rectal cancer patients aged 60 or younger) for a WTP of 25,000 per life-year. That is, at this value, the benefits of surveillance are not worth the costs. If a life-year were valued at \$50,000 or more, optimal follow-up would include routine CEA testing and 1-2 CT scans for all stage III patients 65 and younger and for all considered stage II cohorts, i.e., patients 75 and younger. However, at a WTP just \$5,000 higher (\$55,000), routine CEA testing and at least 1 CT scan would be optimal for all considered cohorts with the exception of 75-year-old stage III colon cancer patients. Including this cohort would require a WTP of \$65,000 or greater.

It is apparent that aggressive use of CT scans multiple years after the primary diagnosis is in general unlikely to be considered a cost-effective use of resources. Unless we value a life-year at \$100,000 or more, annual CT scans for 5 or even 3 years would generally be considered excessive care. Even at a WTP of \$125,000, use of 3 or more CTs would in general only be justified for patients aged 60 or younger, and it would never be optimal for stage III rectal cancer patients (at least among considered ages and WTP values). The main exception to this finding is stage II rectal cancer. At \$75,000, annual CT for 3 years would be optimal among this cohort for patients aged 60 and younger. At \$100,000 (\$125,000), 3-5 CT scans would be optimal for this cohort for patients 65 (70) and younger. I note that younger (≤ 65) stage II rectal cancer patients are the only cohorts for which 5 years of annual CT scan would ever be considered optimal at a WTP value of \$150,000 or less.

Several additional patterns are apparent in Table 14. First, for any combination of disease stage and location and the WTP threshold, the optimal follow-regimen generally

becomes less aggressive with age. Second, the optimal surveillance protocol for stage II and III colon cancer patients and stage III rectal cancer patients was generally the same or very similar, particularly among younger cohorts. If parsimonious recommendations were desired, it would not be unreasonable to treat these three groups as one. Third, among the 3 younger age groups, optimal follow-up for stage II rectal cancer patients was at least as intensive as, and often more so than, that of the other cohorts. However, in oldest age-group, this ordering was not consistent.

If a policy-maker preferred a discount rate of 0%, optimal surveillance intensity would increase, but not dramatically. At a WTP of \$25,000, routine CEA and 1 CT scan would generally only be optimal for patients 60 and younger. At a WTP of \$50,000, optimal follow-up would include CEA and 1-3 CT scans for all cohorts studied. I note again that with 1 exception, the same result was found in the reference-case analysis at a WTP of \$55,000. The main difference would be that at WTP values of \$75,000 and for patients 70 and younger, more aggressive use of CT scans in later years would generally be optimal.

SENSITIVITY ANALYSES

The results of the sensitivity analyses using a discount rate of 3% are presented in Table 15. The undiscounted results are presented in Table D.7 (undiscounted) and are not discussed further. It is apparent that the ICER estimates comparing routine CEA and annual CT scans for 5 (3) years vs. for 3 (2) years were imprecise. This is again due to the extremely small incremental gains in LE associated with these upgrades. However, in the case of 5 years of CT scans at least, it is a moot point given that this aggressive follow-up regimen would not be considered optimal even at a WTP of \$200,000 per life-

year. At a WTP of \$150,000, annual CT scans for more than 2 years would not be considered either.

Among studied cost variables, the costs associated with continuing care for patients after salvage treatment had the biggest influence on the cost-effectiveness results when considered in isolation. Estimates of the ICERs were particularly very sensitive to the choice between using estimates of continuing costs for stage III vs. for stage IV patients. Under the latter assumption, ICERs generally increased by at least 100% and no surveillance was optimal for any of the considered WTP values less than \$100,000. Though at a WTP of \$100,000, the optimal follow-up regimens were similar (routine CEA and 1 CT vs. 2 CTs), and at any higher values they were identical. Using deflated continuing costs for stage III patients, optimal follow-up would involve an additional CT scan at a WTP of \$100,000 or greater but remain unchanged otherwise.

The results were also sensitive to the price of salvage treatment. Using a WTP value of \$50,000, if the cost of salvage treatment was 50% larger (smaller) than assumed in the reference-case analysis, optimal follow-up would change from routine CEA testing and 1 CT scan to no surveillance (CEA and 2 CT scans). However, with higher WTP values, optimal clinical strategies were more robust. With a WTP value of \$75,000, such an increase in the cost of salvage treatment would only mean dropping a second CT, and a 50% decrease in the cost would have no effect on optimal follow-up. In the case of a

WTP of \$100,000, increased salvage costs did not influence optimal treatment, and decreased costs warranted adding a 3rd CT scan.²¹

While the ICER estimates exhibited moderate sensitivity to the cost of palliative treatment, the clinical-decision results were more robust. As predicted, reduced (increased) costs associated with palliative treatment lead to larger (smaller) ICERs. However, using a WTP value of \$50,000 or \$100,000, a 50% increase in the cost of palliative treatment would only lead to the inclusion of an additional CT scan in the optimal regimen. At a WTP value of \$25,000 or \$75,000, there would be no change. A 50% reduction would lead to potentially dropping routine CEA at a WTP of \$50,000, but an even mildly smaller reduction (e.g, 40%) would likely have no effect. Moreover, a 50% reduction in the cost of palliative care would have limited to no effect on the results of the decision analysis at a WTP of \$75,000 or more.

Clinical-decision results were in general robust to univariate variation in studied surveillance-related costs. The only exception to this was the analysis in which the costs of all clinical visits were included and they were assumed to run 60 minutes instead of 30

²¹ The reader may be surprised to see the ICER associated with adding a 3rd CT scan was cheaper in the analyses using increased salvage costs than it was in the reference case. In fact, this happened for several of the analyses. However, this is an artifact of the unstable, miniscule estimates of the discounted incremental effectiveness associated with adding a 3rd CT scan. Mean overall costs were in fact larger in the former analysis (1.5 x salvage costs) than in the reference case as would be expected. The difference is due to a slightly larger estimate of (discounted) incremental LE in the sensitivity analysis. This matters because, when the denominator of a ratio is near 0, slight changes can dominate the effects of larger changes in the numerator. This difference in estimates of the incremental LE is likely due to the use of a smaller Monte Carlo sample size and a failure of the sensitivity analyses to fully sample from the full distribution of time-to-detectable-disease and time-to-clinical-disease parameters.

minutes. At a WTP of \$50,000, optimal surveillance would no longer include routine CEA. However, at other considered WTP values, there would be no change.²²

Table 15: Sensitivity Results: Incremental Cost-Effectiveness Ratios (ICER)

Analysis	1 CT Scan (\$/LY)	CEA + 1 CT Scan (\$/LY)	CEA + 2 CT Scans (\$/LY)	CEA + 3 CT Scans (\$/LY)	CEA + 5 CT Scans (\$/LY)
Discount Rate = 3%					
Baseline	41,061 (40,649-41,473)	46,079 (45,681-46,477)	62,115 (58,895-65,335)	183,271 (142,130-224,412)	258,245 (205,593-310,897)
Salvage Treatment Costs × 1.5	54,481 (53,801-55,161)	59,312 (58,674-59,950)	83,471 (72,268-94,674)	140,705 (92,588-188,822)	515,867 (29,849-1,001,885)
Salvage Treatment Costs × 0.5	24,368 (23,984-24,752)	27,715 (27,392-28,038)	42,888 (37,106-48,670)	86,795 (57,217-116,373)	366,000 (21,323-710,677)
Palliative Costs × 1.5	30,975 (30,401-31,549)	35,878 (35,416-36,340)	47,711 (40,857-54,565)	98,038 (64,118-131,958)	417,367 (23,840-810,894)
Palliative Costs × 0.5	47,878 (47,278-48,478)	51,145 (50,609-51,681)	78,642 (68,156-89,128)	129,474 (85,328-173,620)	464,633 (27,141-902,125)
Salvage Continuing Costs = Stage IV	84,136 (83,203-85,069)	90,046 (89,203-90,889)	122,818 (107,949-137,687)	198,782 (134,838-262,726)	651,467 (49,362-1,253,572)
Salvage Continuing Costs = Stage III × 0.5	29,978 (29,532-30,424)	33,685 (33,283-34,087)	50,572 (43,431-57,713)	95,808 (62,240-129,376)	396,500 (20,557-772,443)
Cost of CT Scan × 2.0	45,642 (45,065-46,219)	45,020 (44,536-45,504)	66,973 (58,023-75,923)	119,538 (78,720-160,356)	461,500 (26,644-896,356)

²² The reader may again note that one or both of the ICERs comparing routine CEA and 1 CT scan to just 1 CT scan and 1 CT scan to no surveillance were unexpectedly less in the sensitivity analyses with inflated surveillance costs (an inflated CT cost and an inflated cost of true and false positives) than they were in the reference-case analysis. Further exploration showed that the incremental surveillance costs were in fact greater in the sensitivity analyses as would be expected. However, this difference was small and was trumped by random variation in incremental LEs and treatment costs due to the sensitivity analyses using fewer iterations and smaller Monte Carlo sizes.

Cost of CT Scan× 0.5	36,770 (36,300-37,240)	42,107 (41,654-42,560)	50,765 (44,111-57,419)	88,821 (59,018-118,624)	325,567 (20,842-630,292)
More Expensive Positive CT Scan Results	39,998 (39,490-40,506)	44,299 (43,821-44,777)	64,166 (55,594-72,738)	115,692 (76,198-155,186)	449,200 (25,961-872,439)
More Expensive Clinical Visits	41,954 (41,422-42,486)	57,732 (57,097-58,367)	63,032 (54,623-71,441)	113,667 (74,903-152,431)	443,233 (25,706-860,760)
5-FU for Unresectable Recurrence	43,215 (42,785-43,645)	46,712 (46,316-47,108)	61,355 (55,181-67,529)	95,911 (73,253-118,569)	233,077 (132,706-333,448)
Salvage Mortality Rate = 5%	40,907 (40,349-41,465)	43,206 (42,723-43,689)	68,300 (57,790-78,810)	144,271 (79,553-208,989)	281,234 (116,665-445,803)

Table 15 gives the results of the sensitivity analyses using a discount rate of 3%. ICERs are reported in 2017 US dollars per additional life-year. The first row (Baseline) gives the reference-case ICERS for 65-year-old stage III colon cancer patients. Each subsequent row gives the ICERS (and 95% CIs) for different univariate sensitivity analyses. The first four involved multiplying the assumed cost(s) by a factor of 1.5 or 0.5. The 5th row in each section reports the results when the continuing costs among patients treated with salvage treatment were assumed to be the same as for patients with unresectable recurrence. For the latter patients, the model used an estimate based on stage IV patients. In the reference-case analysis, the model used an estimate for continuing costs following salvage treatment based on stage III patients. The row immediately below reports the results using continuing costs equal to 50% of those of stage III patients. The next two rows present the results using an inflated and deflated cost for a CT scan, respectively. In the following two rows, analyses used larger values for the cost of a true-positive (x10) and false-positive (2x) CT scan result and an increased cost of routine clinical visits, respectively. Finally, the last two rows give the results for an analysis done assuming all (65-year old stage III colon cancer) patients with unresectable recurrence are treated with 5-FU rather than FOLFOX and for an analysis done using a salvage mortality rate of 5% rather than 2%. To determine the impact variation in a given parameter had on the ICERs, the ICERs from the appropriate row should be compared to the baseline row.

The results were also robust to the two efficacy parameters studied. First, the conclusions of the decision analysis for 65-year-old stage III colon cancer patients would not change for any WTP value of \$75,000 or less if all patients with unresectable

recurrence in that cohort were treated with a single-agent systemic chemotherapy (5-FU) rather than a double-agent regimen like FOLFOX. For a WTP of \$100,000 or greater, an additional (3rd) CT scan would be optimal under this scenario. Second, the decision-analysis was also robust to an increase in the 30-day mortality rate associated with salvage surgery from 2% to 5%. In fact, ICER point estimates actually slightly decreased for two comparisons, and for at least one of the comparisons this difference was not due to Monte Carlo imprecision. Further exploration showed that, while the incremental LE was smaller under this assumption, the incremental costs were also smaller, and the latter difference was more important than the former. In particular, the difference in incremental costs was driven by a reduction in the costs of treatment, presumably due to the absence of the continuing costs.

DISCUSSION

The results presented above suggest that, if salvage surgery induced by extra-colonic follow-up does confer a survival advantage as predicted by the model, moderately intensive surveillance may be cost-effective for stage II and III colon and rectal cancer patients, depending upon the value we place on a life-year and the age of the patients. At a WTP threshold of \$25,000, resources would be better spent on other health-related interventions. However, at a WTP of \$55,000 or more, routine CEA testing and at least one or two CT scans should be offered to all patients aged 75 or younger, with the possible exception of 75-year old stage III colon-cancer patients.

More generally, if we decided our WTP for a life-year was greater than \$50,000 but less than \$100,000, a reasonable and simple guideline would be as follows. All locally-advanced CRC patients aged 75 or younger who were initially treated with curative

resection and are otherwise healthy (i.e., would be candidates for salvage surgery) should be offered routine CEA testing in accordance with current US guidelines (every 3 months for 2 years and then every 6 months for 3 years using a threshold of 5 µg/L) along with a chest and abdominal (abdominal-pelvic for rectal patients) CT scan at 12-18 months. Moreover, providers may consider suggesting annual CT scans for 2 years (instead of just 1) for patients who either (a) present with stage II rectal cancer or (b) are aged 60 or younger (and possibly for those aged 61-65). If a patient presented with both of these features, a 3rd scan at 3 years might also be considered.

For a WTP threshold between \$100,000 and \$150,000, a reasonable recommendation would be as follows. All patients aged 70 or younger who were initially treated with curative resection and are otherwise healthy should be offered routine CEA testing (as above) and annual CT scans for at least 2 years. For patients (a) with stage II rectal cancer or (b) who are 60 or younger, a 3rd CT scan at 3 years could be offered. Moreover, annual CT scans for up to 5 years might be considered for stage II rectal cancer patients aged 60 or younger. Finally, patients between the ages of 71 and 75 should be offered routine CEA testing and a single CT scan at 12-18 months.

While univariate sensitivity analyses are inadequate to assess the full impact of parameter uncertainty on results of the cost-effectiveness analysis and to quantify decision uncertainty¹⁶², they can determine whether a particularly uncertain parameter or cost is likely to have a large effect on the results. The results presented above identified one clearly very important uncertain set of costs – continuing costs after salvage therapy - and one set of uncertain cost variables of potentially modest importance – the upfront cost of salvage treatment.

The values used for monthly continuing costs after salvage therapy are important because they accrue for a long period of time. A patient who survives 50 months after salvage therapy (slightly less than the median survival) but ultimately dies from cancer would accrue 35 months of continuing costs. Since this patient would likely be under 75 for the entire period, this would amount to over \$35,000 just in continuing costs if the rate for stage III patients was used. At the rate assigned to stage IV patients, this would amount to about \$125,000. I note again that this amount is over and above the cost of any second salvage attempt.

I was unable to find a source in the literature which directly estimated continuing costs for patients after salvage therapy for recurrence generally or for just isolated metastatic disease. However, it seems improbable that there would have been many such patients in the SEER-Medicare cohort (incidence cases from 1996-2002) that was used to estimate continuing costs for stage IV patients, a cohort which had less than 10% survival at 5 years. Thus, there is reason to suspect that using the stage IV monthly costs would overestimate the true continuing costs that accrue to patients who undergo a clear-margin surgical resection of recurrence. I note that the estimate of stage III monthly continuing costs for patients under 75 used in this paper is very similar to the estimate of such costs used in Gazelle, et al., (2003) once adjusted for inflation.¹⁵³ Still, further research into the true value of these continuing costs should be welcomed as it would help reduce significant uncertainty surrounding the value of ICERs and could even lead to a change in recommendations. Though at WTP threshold of \$100,000 or greater, decision-uncertainty attributable to uncertainty surrounding continuing costs is likely to be small.

The mean cost of salvage therapy was less important to the clinical-decision results in a univariate context than were continuing costs, but it was still an important and uncertain cost variable. Of particular concern is whether the costs are notably larger than the values used in the reference-case analysis. At a WTP of \$75,000 or greater, this seems to be less of a concern, but at a WTP threshold of \$50,000 it is plausible a significant increase in the cost of salvage treatment could render no surveillance optimal. While my estimates were based mostly on Medicare reimbursement rates, these estimates inevitably only constitute informed guesses at what an average cost would be for salvage treatment of recurrence for rectal and colon cancer patients. They are uncertain because of all the assumptions (e.g., utilization, types of procedures, the risk of morbidity, the cost associated with mortality, etc) that were required to construct them. While I could not find any other such estimates in the literature, Gazelle et al., (2003) reported an estimate of the mean cost associated with a completed hepatic resection of about \$54,000 (inflated to 2017 US dollars). This estimate included the expected costs associated with serious morbidity and mortality. This estimate is very similar to my estimate of the expected cost of hepatic salvage treatment at about \$52,000 (Table D.3). However, unlike the estimate used in this paper, the former estimate does not include costs associated with (neo)adjuvant chemotherapy because the authors did not include any such costs. Even though both papers used Medicare reimbursement rates for physician and inpatient services, my estimate of the mean surgery-related cost was notably smaller, about \$35,000. This large difference seems to be the result of a significant decline in the real value of relevant Medicare inpatient and physician reimbursement rates. I note that if the same costs among patients with private insurance have kept pace with the inflation of

medical prices instead of declining in real terms, the sensitivity analysis using a value of 1.5 times the reference-case estimate for the price of salvage treatment might be more relevant for such patients.

An advantage of recommending at least routine CEA testing and a CT scan at 12-18 months is that it would realize the overwhelming majority of the predicted survival benefit attributable to intensive extra-colonic follow-up. For example, for 65-year-old stage III colon cancer patients, 91% of the total discounted gains in LE attributable to a surveillance schedule including routine CEA testing and annual CT for 5 years (relative to no surveillance) was realized by routine CEA testing and a single CT scan at 12-18 months. Almost a third of that benefit could be realized by just offering a single CT scan. More skeptical policymakers or panel members might consider hedging their bet and only recommending this surveillance strategy of modest intensity but not further use of CT scans, even for younger or stage II rectal cancer patients. The loss to patients from denying or dissuading further follow-up (i.e., additional CT scans) would be very small even if it proved to be technically efficacious (as the model predicts). On the other hand, if the causal theory embodied in the model is correct and policymakers or panel-members discourage routine CEA testing and a single CT scan because of lack of experimental evidence, the forgone public-health benefits would be quite a bit larger. To help put this in perspective, consider that the predicted (discounted) incremental gain in LE (73 days) attributable to 65-year-old stage III colon cancer patients undergoing routine CEA testing and a single CT scan (vs. no follow-up) is over 3 times larger than the similarly discounted predicted gain in LE of 23 days attributable to moving from no CRC screening to the most effective regimen identified among 22 possible strategies

considered in a cost-effectiveness analysis of CRC screening among average-risk 50 year olds.¹⁷⁵ This discrepancy is partially driven by the assumption of 100% compliance in these analyses and only 60% in the cited report, but clearly this is a health benefit most would consider to be of importance. Moreover, over 80% of the incremental costs associated with this strategy compared to no surveillance would be attributable to treatment cost rather than surveillance costs. While this fact is strictly speaking irrelevant to the determination of the optimal strategy, it may make the strategy more palatable to the policymakers or panel members.

The strategy of routine CEA testing and a single CT scan at 12-18 months is identical to the recommendation made by the FACS trial investigators who used the trial data to conduct a cost-utility analysis from the perspective of the UK NHS using a time-horizon of 5 years.⁴⁵ In addition, the finding that stage II rectal cancer patients should be offered at least as intensive and generally more intensive extra-colonic surveillance as is offered to the other patients is consistent with the results of two other RCTs evaluating intensive surveillance. The trial that reported the largest mortality reduction (and was one of two trials to find a statistically significant benefit) appears to have enrolled only rectal cancer patients.⁴¹ Moreover, a relatively more recent imaging trial conducted in Spain found no survival benefit from intensifying follow-up overall, but a subgroup analysis found a significant benefit among stage II patients and rectal cancer patients.¹⁴⁶ Finally, the authors of the FACs trial reported that a similar proportion of enrolled stage III and stage II patients underwent salvage surgery.⁴⁶ That is, even though the recurrence rate was greater among stage III patients, the increased salvage rate among stage II patients adequately compensated for this.

One limitation of the microsimulation model's inevitably simplified representation of disease development and detection is that it does not distinguish among different anatomical locations of recurrence. For example, the model does not distinguish between pelvic recurrence or hepatic or pulmonary metastases. This is potentially relevant for evaluating surveillance among rectal cancer patients because CEA tests appear to have a lower sensitivity with local-regional recurrence and pulmonary metastases^{23 26} and rectal cancer patients are more likely to suffer these types of recurrence (and undergo salvage treatment for them) than are colon cancer patients. Therefore, it seems likely that if this simplification biases the results it would be in the direction of underestimating the marginal value of CT scans for rectal cancer patients. Thus, less conservative policymakers or panel members might consider recommending a 2nd CT scan for all rectal cancer patients (or a 3rd CT scan in the recommendation based on a higher range of WTP values).

Both the above consideration and the results of this paper suggest that oncologists might consider more aggressively encouraging appropriate stage II patients (particularly stage II rectal cancer patients) to comply at least with a CT scan at 12-18 months and routine CEA testing as this group is apparently less prone to comply with suggested surveillance.¹⁷⁶ This is presumably because of the incorrect belief communicated by physicians, or at least implicitly transmitted, to patients that stage II patients will benefit much less on average from such surveillance due to their lower recurrence rates.

Another potential limitation of this paper is that I assumed 100% compliance with surveillance tests. In this sense, the cost-effectiveness analyses are based on the expected maximum potential benefit. One reason I made this assumption was to increase

computational efficiency. As noted above, the discounted incremental gains in LE associated with later CT scans were extremely small when averaged over the entire cohort. This tended to lead to imprecise ICER estimates among some cohorts and comparisons even with a Monte Carlo size of 75 million per arm. Including poor compliance in the analyses would have compounded this problem given that the benefits would now be limited to a smaller population. Still, 100% compliance is obviously not a realistic assumption and so the consequences should be considered.¹⁷⁶⁻¹⁷⁸ Poor adherence will undoubtedly lead to a reduction in the impact of recommended surveillance. However it is less clear that it will impact the cost-effectiveness of alternative strategies and the selection of optimal care. Compliance has been a hot topic in recent analyses of alternative CRC screening strategies and for good reason.^{179 180} There is reportedly significant variation in patient compliance depending on the test, e.g., sigmoidoscopy, colonoscopy, or CT colonography. In the case of post-diagnosis surveillance, however, it is less clear that this is an issue. Patients who stop attending routine clinical visits and undergoing CEA testing are also likely to stop getting a CT scan and vice versa. These patients would fail to get the residual benefit from further surveillance testing, but they would also stop using additional resources associated with surveillance and surveillance-induced salvage therapy. These patients essentially opt into the baseline treatment arm of no surveillance. The most plausible way poor adherence could affect the results of the cost-effectiveness analysis would be if patients stopped undergoing CEA testing, say after 1-2 years, but would be willing to undergo a second CT scan near the end of that period. Additional empirical research on compliance in the Medicare population in

contemporary practice would help evaluate this possibility. Still, future research should formally investigate such possibilities.

The recommendations offered above inevitably are limited by the set of interventions studied in this paper. For example, in this paper, I did not investigate more intensive CT imaging concentrated in the first year or two as has been studied in the FACS and GILDA trials. I omitted such strategies because they were not included in any of the most recent professional-society guidelines the way that varying lengths of annual CT scans were. Moreover, given the already computationally burdensome nature of the analyses of this paper, it would have been costly to increase the number of comparisons. I instead chose to prioritize performing separate analyses for stage, location, and age. Future research could focus on optimizing the timing and frequency of CT scans occurring in the first 24 months.

Similarly, the above recommendations do not apply to stage I patients or patients older than 75. It is unlikely that intensive CT imaging would be cost-effective in either group, but future research should investigate these questions. For example, it is possible that somewhat regular CEA testing and a CT scan at 12-18 months might be appropriate for stage I rectal cancer patients. In the cast of patients 80 or older, it is unclear that many of the current model parameters would apply to this population.

The above recommendations accept the causal hypothesis of the model. However, the above results obviously do not establish the truth of that hypothesis. In Chapter 3, I argued in support of this theory, but clearly it is supported by weak evidence. As noted in the introduction, policymakers and payers thus face a decision in the absence of good evidence. In the abstract, this decision problem involves making a choice among multiple

mutually-exclusive clinical strategies (one of which is offering no extra-colonic follow-up) with imperfect information about their relative benefits (among other things). In principle, a formal decision analysis could be used to resolve this dilemma and determine whether further information should be collected. While such an analysis obviously cannot create additional evidence, it can be used to identify the optimal strategy given the available evidence. In a context where the goal is to maximize health gains within a given budget, it has been convincingly argued that the optimal (clinical) strategy is that with the highest expected net benefit, regardless of the results of any statistical inference.¹⁸¹ The idea then would be to make a probabilistic sensitivity analysis (PSA) the main analysis (rather than a sensitivity analysis) and include our full uncertainty about the benefit of intensive surveillance in that analysis. We would then select the strategy with the greatest mean NMB as optimal (which might be no surveillance) given the evidence we have and secondly evaluate whether the expected benefits of gathering further information, e.g., a very large RCT, would outweigh the expected costs. Decision scientists might wonder why such an approach was not taken.

Crucially, such an analysis would require formalizing the full extent of our uncertainty concerning the benefit of intensive follow-up. A simple approach would be to use the meta-analytic predictive distribution of the all-cause mortality hazard ratio (HR). However, this approach would be problematic for two reasons. First, the meta-analyses include multiple studies performed in an era in which patients faced a notably higher recurrence rate than they do in contemporary practice. I showed in Chapter 3 of this dissertation that the recurrence rate is an important determinant of the magnitude of the reduction in mortality hazards associated with surveillance. Second, any meta-analysis

would inevitably include a heterogeneous set of interventions grouped as ‘intensive’, ‘minimal’, or no surveillance. However, we are really interested in more discriminating comparisons, e.g., comparing routine CEA and 1 CT scan to 2 or 3 CT scans. A related problem is that the model was constructed in such a way that there is no such single parameter governing the benefit of surveillance. Rather, because we are interested in the latter type of comparisons, the model explicitly characterizes the development, detection, and treatment of recurrence and differing mortality experiences after different treatments for recurrence.

In the current implementation, the benefit of salvage therapy basically consists of the discrepancy between two survival curves: one for patients treated with an R0 resection and one for patients for which curative resection is not possible. As described in Chapter 2, the latter were taken from RCTs of systemic chemotherapy for patients with incurable disease. The former set of survival curves are based upon those observed in a large community-practice case-series. However, as noted in Chapter 3, they are consistent with the survival experience observed among patients treated with curative intent in the FACS trial and in other unselected cohort studies involving intensive follow-up. Thus, a full decision analysis would require implicitly or explicitly putting a distribution on this discrepancy. While technically feasible, it is unclear what sort of distribution would be appropriate and non-arbitrary. While it is worth exploring this option in future research, it is likely that the results of the analysis would depend heavily on the distribution chosen.

Bibliography

1. Siegel RL, Miller KD, Fedewa SA, et al. Colorectal cancer statistics, 2017. *CA: a cancer journal for clinicians* 2017;67(3):177-93. doi: 10.3322/caac.21395 [published Online First: 2017/03/02]
2. Bouvier AM, Launoy G, Bouvier V, et al. Incidence and patterns of late recurrences in colon cancer patients. *International journal of cancer* 2015;137(9):2133-8. doi: 10.1002/ijc.29578 [published Online First: 2015/04/29]
3. Sargent D, Sobrero A, Grothey A, et al. Evidence for cure by adjuvant therapy in colon cancer: observations based on individual patient data from 20,898 patients on 18 randomized trials. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 2009;27(6):872-7. doi: 10.1200/jco.2008.19.5362 [published Online First: 2009/01/07]
4. Augestad KM, Bakaki PM, Rose J, et al. Metastatic spread pattern after curative colorectal cancer surgery. A retrospective, longitudinal analysis. *Cancer epidemiology* 2015;39(5):734-44. doi: 10.1016/j.canep.2015.07.009 [published Online First: 2015/08/19]
5. Sjøvall A, Granath F, Cedermark B, et al. Loco-regional recurrence from colon cancer: a population-based study. *Annals of surgical oncology* 2007;14(2):432-40. doi: 10.1245/s10434-006-9243-1 [published Online First: 2006/12/02]
6. Renehan AG, Egger M, Saunders MP, et al. Impact on survival of intensive follow up after curative resection for colorectal cancer: systematic review and meta-analysis of randomised trials. *BMJ : British Medical Journal* 2002;324(7341):813-13.
7. Pickhardt PJ, Edwards K, Bruining DH, et al. Prospective Trial Evaluating the Surgical Anastomosis at One-Year Colorectal Cancer Surveillance: CT Colonography Versus Optical Colonoscopy and Implications for Patient Care. *Diseases of the colon and rectum* 2017;60(11):1162-67. doi: 10.1097/dcr.0000000000000845 [published Online First: 2017/10/11]
8. Rahbari NN, Ulrich AB, Bruckner T, et al. Surgery for locally recurrent rectal cancer in the era of total mesorectal excision: is there still a chance for cure? *Annals of surgery* 2011;253(3):522-33. doi: 10.1097/SLA.0b013e3182096d4f [published Online First: 2011/01/07]
9. van Gijn W, Marijnen CA, Nagtegaal ID, et al. Preoperative radiotherapy combined with total mesorectal excision for resectable rectal cancer: 12-year follow-up of the multicentre, randomised controlled TME trial. *The Lancet Oncology* 2011;12(6):575-82. doi: 10.1016/s1470-2045(11)70097-3 [published Online First: 2011/05/21]
10. Hellinger MD, Santiago CA. Reoperation for recurrent colorectal cancer. *Clinics in colon and rectal surgery* 2006;19(4):228-36. doi: 10.1055/s-2006-956445 [published Online First: 2006/11/01]
11. Tepper JE, O'Connell M, Hollis D, et al. Analysis of surgical salvage after failure of primary therapy in rectal cancer: results from Intergroup Study 0114. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 2003;21(19):3623-8. doi: 10.1200/jco.2003.03.018 [published Online First: 2003/09/27]
12. Lin Koo S, Wen JH, Hillmer A, et al. Current and emerging surveillance strategies to expand the window of opportunity for curative treatment after surgery in colorectal cancer. *Expert Review of Anticancer Therapy* 2013;13(4):439-50. doi: 10.1586/era.13.14

13. Meyerhardt JA, Mangu PB, Flynn PJ, et al. Follow-up care, surveillance protocol, and secondary prevention measures for survivors of colorectal cancer: American Society of Clinical Oncology clinical practice guideline endorsement. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 2013;31(35):4465-70. doi: 10.1200/jco.2013.50.7442 [published Online First: 2013/11/14]
14. Steele SR, Chang GJ, Hendren S, et al. Practice Guideline for the Surveillance of Patients After Curative Treatment of Colon and Rectal Cancer. *Diseases of the colon and rectum* 2015;58(8):713-25. doi: 10.1097/dcr.0000000000000410 [published Online First: 2015/07/15]
15. El-Shami K, Oeffinger KC, Erb NL, et al. American Cancer Society Colorectal Cancer Survivorship Care Guidelines. *CA: a cancer journal for clinicians* 2015;65(6):428-55. doi: 10.3322/caac.21286 [published Online First: 2015/09/09]
16. Schmoll HJ, Van Cutsem E, Stein A, et al. ESMO Consensus Guidelines for management of patients with colon and rectal cancer. a personalized approach to clinical decision making. *Annals of oncology : official journal of the European Society for Medical Oncology* 2012;23(10):2479-516. doi: 10.1093/annonc/mds236 [published Online First: 2012/09/27]
17. Poston GJ, Tait D, O'Connell S, et al. Diagnosis and management of colorectal cancer: summary of NICE guidance. *BMJ (Clinical research ed)* 2011;343:d6751. doi: 10.1136/bmj.d6751 [published Online First: 2011/11/15]
18. Sargent DJ, Patiyil S, Yothers G, et al. End points for colon cancer adjuvant trials: observations and recommendations based on individual patient data from 20,898 patients enrolled onto 18 randomized trials from the ACCENT Group. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 2007;25(29):4569-74. doi: 10.1200/jco.2006.10.4323 [published Online First: 2007/09/19]
19. Shah MA, Renfro LA, Allegra CJ, et al. Impact of Patient Factors on Recurrence Risk and Time Dependency of Oxaliplatin Benefit in Patients With Colon Cancer: Analysis From Modern-Era Adjuvant Studies in the Adjuvant Colon Cancer End Points (ACCENT) Database. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 2016;34(8):843-53. doi: 10.1200/jco.2015.63.0558 [published Online First: 2016/01/27]
20. Teloken PE, Ransom D, Faragher I, et al. Recurrence in patients with stage I colorectal cancer. *ANZ journal of surgery* 2016;86(1-2):49-53. doi: 10.1111/ans.13254 [published Online First: 2015/08/04]
21. Kobayashi H, Mochizuki H, Sugihara K, et al. Characteristics of recurrence and surveillance tools after curative resection for colorectal cancer: a multicenter study. *Surgery* 2007;141(1):67-75. doi: 10.1016/j.surg.2006.07.020 [published Online First: 2006/12/26]
22. Wichmann MW, Lau-Werner U, Muller C, et al. Carcinoembryonic antigen for the detection of recurrent disease following curative resection of colorectal cancer. *Anticancer research* 2000;20(6d):4953-5. [published Online First: 2001/05/01]
23. Shinkins B, Nicholson BD, Primrose J, et al. The diagnostic accuracy of a single CEA blood test in detecting colorectal cancer recurrence: Results from the FACS trial. *PLoS ONE* 2017;12(3):e0171810. doi: 10.1371/journal.pone.0171810
24. Treasure T, Monson K, Fiorentino F, et al. The CEA Second-Look Trial: a randomised controlled trial of carcinoembryonic antigen prompted reoperation for recurrent

- colorectal cancer. *BMJ open* 2014;4(5):e004385. doi: 10.1136/bmjopen-2013-004385 [published Online First: 2014/05/16]
25. Shinkins B, Nicholson BD, James T, et al. What carcinoembryonic antigen level should trigger further investigation during colorectal cancer follow-up? A systematic review and secondary analysis of a randomised controlled trial. *Health technology assessment (Winchester, England)* 2017;21(22):1-60. doi: 10.3310/hta21220 [published Online First: 2017/06/16]
 26. Goldstein MJ, Mitchell EP. Carcinoembryonic antigen in the staging and follow-up of patients with colorectal cancer. *Cancer investigation* 2005;23(4):338-51. [published Online First: 2005/08/17]
 27. Nicholson BD, Shinkins B, Pathiraja I, et al. Blood CEA levels for detecting recurrent colorectal cancer. *The Cochrane database of systematic reviews* 2015(12):Cd011134. doi: 10.1002/14651858.CD011134.pub2 [published Online First: 2015/12/15]
 28. Sobhani I, Tiret E, Lebtahi R, et al. Early detection of recurrence by 18FDG-PET in the follow-up of patients with colorectal cancer. *British journal of cancer* 2008;98(5):875-80. doi: 10.1038/sj.bjc.6604263 [published Online First: 2008/02/28]
 29. Gonzalez M, Poncet A, Combescure C, et al. Risk factors for survival after lung metastasectomy in colorectal cancer patients: a systematic review and meta-analysis. *Annals of surgical oncology* 2013;20(2):572-9. doi: 10.1245/s10434-012-2726-3 [published Online First: 2012/10/30]
 30. Kahi CJ, Boland CR, Dominitz JA, et al. Colonoscopy Surveillance After Colorectal Cancer Resection: Recommendations of the US Multi-Society Task Force on Colorectal Cancer. *Gastroenterology* 2016;150(3):758-68.e11. doi: 10.1053/j.gastro.2016.01.001 [published Online First: 2016/02/20]
 31. Makela JT, Laitinen SO, Kairaluoma MI. Five-year follow-up after radical surgery for colorectal cancer. Results of a prospective randomized trial. *Archives of surgery (Chicago, Ill : 1960)* 1995;130(10):1062-7. [published Online First: 1995/10/01]
 32. Ohlsson B, Breland U, Ekberg H, et al. Follow-up after curative surgery for colorectal carcinoma. Randomized comparison with no follow-up. *Diseases of the colon and rectum* 1995;38(6):619-26. [published Online First: 1995/06/01]
 33. Pietra N, Sarli L, Costi R, et al. Role of follow-up in management of local recurrences of colorectal cancer: a prospective, randomized study. *Diseases of the colon and rectum* 1998;41(9):1127-33. [published Online First: 1998/09/28]
 34. Schoemaker D, Black R, Giles L, et al. Yearly colonoscopy, liver CT, and chest radiography do not influence 5-year survival of colorectal cancer patients. *Gastroenterology* 1998;114(1):7-14. [published Online First: 1998/01/15]
 35. Rosati G, Ambrosini G, Barni S, et al. A randomized trial of intensive versus minimal surveillance of patients with resected Dukes B2-C colorectal carcinoma. *Annals of oncology : official journal of the European Society for Medical Oncology* 2016;27(2):274-80. doi: 10.1093/annonc/mdv541 [published Online First: 2015/11/19]
 36. Figueredo A, Rumble RB, Maroun J, et al. Follow-up of patients with curatively resected colorectal cancer: a practice guideline. *BMC cancer* 2003;3:26. doi: 10.1186/1471-2407-3-26 [published Online First: 2003/10/08]
 37. Pita-Fernandez S, Alhayek-Ai M, Gonzalez-Martin C, et al. Intensive follow-up strategies improve outcomes in nonmetastatic colorectal cancer patients after curative surgery: a systematic review and meta-analysis. *Annals of oncology : official journal of the*

- European Society for Medical Oncology* 2015;26(4):644-56. doi: 10.1093/annonc/mdu543 [published Online First: 2014/11/21]
38. Rosen M, Chan L, Beart RW, et al. Follow-up of colorectal cancer. *Diseases of the Colon & Rectum* 1998;41(9):1116-26. doi: 10.1007/BF02239433
 39. Tjandra JJ, Chan MK. Follow-up after curative resection of colorectal cancer: a meta-analysis. *Diseases of the colon and rectum* 2007;50(11):1783-99. doi: 10.1007/s10350-007-9030-5 [published Online First: 2007/09/18]
 40. Jeffery M, Hickey BE, Hider PN. Follow-up strategies for patients treated for non-metastatic colorectal cancer. *The Cochrane database of systematic reviews* 2007(1):Cd002200. doi: 10.1002/14651858.CD002200.pub2 [published Online First: 2007/01/27]
 41. Secco GB, Fardelli R, Gianquinto D, et al. Efficacy and cost of risk-adapted follow-up in patients after colorectal cancer surgery: a prospective, randomized and controlled trial. *European journal of surgical oncology : the journal of the European Society of Surgical Oncology and the British Association of Surgical Oncology* 2002;28(4):418-23. [published Online First: 2002/07/09]
 42. Parmar MK, Torri V, Stewart L. Extracting summary statistics to perform meta-analyses of the published literature for survival endpoints. *Statistics in medicine* 1998;17(24):2815-34. [published Online First: 1999/01/28]
 43. Tierney JF, Stewart LA, Ghersi D, et al. Practical methods for incorporating summary time-to-event data into meta-analysis. *Trials* 2007;8:16. doi: 10.1186/1745-6215-8-16 [published Online First: 2007/06/09]
 44. Rose J, Augestad KM, Cooper GS. Colorectal cancer surveillance: what's new and what's next. *World journal of gastroenterology* 2014;20(8):1887-97. doi: 10.3748/wjg.v20.i8.1887 [published Online First: 2014/03/04]
 45. Mant D, Gray A, Pugh S, et al. A randomised controlled trial to assess the cost-effectiveness of intensive versus no scheduled follow-up in patients who have undergone resection for colorectal cancer with curative intent. *Health technology assessment (Winchester, England)* 2017;21(32):1-86. doi: 10.3310/hta21320 [published Online First: 2017/06/24]
 46. Primrose JN, Perera R, Gray A, et al. Effect of 3 to 5 years of scheduled CEA and CT follow-up to detect recurrence of colorectal cancer: the FACS randomized clinical trial. *Jama* 2014;311(3):263-70. doi: 10.1001/jama.2013.285718 [published Online First: 2014/01/17]
 47. Jeffery M, Hickey BE, Hider PN, et al. Follow-up strategies for patients treated for non-metastatic colorectal cancer. *The Cochrane database of systematic reviews* 2016;11:Cd002200. doi: 10.1002/14651858.CD002200.pub3 [published Online First: 2016/11/25]
 48. Mokhles S, Macbeth F, Farewell V, et al. Meta-analysis of colorectal cancer follow-up after potentially curative resection. *The British journal of surgery* 2016;103(10):1259-68. doi: 10.1002/bjs.10233 [published Online First: 2016/08/05]
 49. Treasure T, Russell C, Macbeth F. Re-launch of PulMiCC trial to discover the true effect of pulmonary metastasectomy on survival in advanced colorectal cancer. *BMJ : British Medical Journal* 2015;351 doi: 10.1136/bmj.h6045
 50. Treasure T, Macbeth F. Is Surgery Warranted for Oligometastatic Disease? *Thoracic surgery clinics* 2016;26(1):79-90. doi: 10.1016/j.thorsurg.2015.09.010 [published Online First: 2015/11/28]

51. Åberg T, Treasure T. Analysis of pulmonary metastasis as an indication for operation: an evidence-based approach. *European Journal of Cardio-Thoracic Surgery* 2016;50(5):792-98. doi: 10.1093/ejcts/ezw140
52. Treasure T, Macbeth F. The GILDA trial finds no survival benefit from intensified screening after primary resection of colorectal cancer: the PulMiCC trial tests the survival benefit of pulmonary metastasectomy for detected asymptomatic lung metastases. *Annals of Oncology* 2016;27(4):745-45. doi: 10.1093/annonc/mdv618
53. Kumar R, Price TJ, Beeke C, et al. Colorectal cancer survival: An analysis of patients with metastatic disease synchronous and metachronous with the primary tumor. *Clinical colorectal cancer* 2014;13(2):87-93. doi: 10.1016/j.clcc.2013.11.008 [published Online First: 2014/01/01]
54. Broadbridge VT, Karapetis CS, Beeke C, et al. Do metastatic colorectal cancer patients who present with late relapse after curative surgery have a better survival? *British journal of cancer* 2013;109(5):1338-43. doi: 10.1038/bjc.2013.388 [published Online First: 2013/07/19]
55. Kobayashi H, Mochizuki H, Morita T, et al. Timing of relapse and outcome after curative resection for colorectal cancer: a Japanese multicenter study. *Digestive surgery* 2009;26(3):249-55. doi: 10.1159/000226868 [published Online First: 2009/07/03]
56. O'Connell MJ, Campbell ME, Goldberg RM, et al. Survival following recurrence in stage II and III colon cancer: findings from the ACCENT data set. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 2008;26(14):2336-41. doi: 10.1200/jco.2007.15.8261 [published Online First: 2008/05/10]
57. Rutter CM, Johnson EA, Feuer EJ, et al. Secular Trends in Colon and Rectal Cancer Relative Survival. *JNCI Journal of the National Cancer Institute* 2013;105(23):1806-13. doi: 10.1093/jnci/djt299
58. Shi Q, Andre T, Grothey A, et al. Comparison of outcomes after fluorouracil-based adjuvant therapy for stages II and III colon cancer between 1978 to 1995 and 1996 to 2007: evidence of stage migration from the ACCENT database. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 2013;31(29):3656-63. doi: 10.1200/jco.2013.49.4344 [published Online First: 2013/08/28]
59. Miller KD, Siegel RL, Lin CC, et al. Cancer treatment and survivorship statistics, 2016. *CA: a cancer journal for clinicians* 2016;66(4):271-89. doi: 10.3322/caac.21349 [published Online First: 2016/06/03]
60. Siegel RL, Fedewa SA, Anderson WF, et al. Colorectal Cancer Incidence Patterns in the United States, 1974–2013. *JNCI: Journal of the National Cancer Institute* 2017;109(8):djw322-djw22. doi: 10.1093/jnci/djw322
61. Tanis PJ, Doeksen A, van Lanschot JJ. Intentionally curative treatment of locally recurrent rectal cancer: a systematic review. *Canadian journal of surgery Journal canadien de chirurgie* 2013;56(2):135-44. doi: 10.1503/cjs.025911 [published Online First: 2013/03/23]
62. Lindholm E, Brevinge H, Haglund E. Survival benefit in a randomized clinical trial of faecal occult blood screening for colorectal cancer. *The British journal of surgery* 2008;95(8):1029-36. doi: 10.1002/bjs.6136 [published Online First: 2008/06/20]
63. Lin JS, Piper MA, Perdue LA, et al. U.S. Preventive Services Task Force Evidence Syntheses, formerly Systematic Evidence Reviews. Screening for Colorectal Cancer: A Systematic

- Review for the US Preventive Services Task Force. Rockville (MD): Agency for Healthcare Research and Quality (US) 2016.
64. R: A language and environment for statistical computing. [program]. Vienna, Austria: R Foundation for Statistical Computing, 2013.
 65. Booth CM, Nanji S, Wei X, et al. Surgical resection and peri-operative chemotherapy for colorectal cancer liver metastases: A population-based study. *European journal of surgical oncology : the journal of the European Society of Surgical Oncology and the British Association of Surgical Oncology* 2016;42(2):281-7. doi: 10.1016/j.ejso.2015.10.006 [published Online First: 2015/11/13]
 66. Booth CM, Nanji S, Wei X, et al. Outcomes of Resected Colorectal Cancer Lung Metastases in Routine Clinical Practice: A Population-Based Study. *Annals of surgical oncology* 2016;23(4):1057-63. doi: 10.1245/s10434-015-4979-0 [published Online First: 2015/11/18]
 67. Harris CA, Solomon MJ, Heriot AG, et al. The Outcomes and Patterns of Treatment Failure After Surgery for Locally Recurrent Rectal Cancer. *Annals of surgery* 2016;264(2):323-9. doi: 10.1097/sla.0000000000001524 [published Online First: 2015/12/23]
 68. Karnon J, Stahl J, Brennan A, et al. Modeling using discrete event simulation: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force--4. *Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research* 2012;15(6):821-7. doi: 10.1016/j.jval.2012.04.013 [published Online First: 2012/09/25]
 69. Caro JJ, Moller J, Getsios D. Discrete event simulation: the preferred technique for health economic evaluations? *Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research* 2010;13(8):1056-60. doi: 10.1111/j.1524-4733.2010.00775.x [published Online First: 2010/09/10]
 70. Othus M, Barlogie B, Leblanc ML, et al. Cure models as a useful statistical tool for analyzing survival. *Clinical cancer research : an official journal of the American Association for Cancer Research* 2012;18(14):3731-6. doi: 10.1158/1078-0432.ccr-11-2859 [published Online First: 2012/06/08]
 71. Seo SI, Lim SB, Yoon YS, et al. Comparison of recurrence patterns between ≤ 5 years and > 5 years after curative operations in colorectal cancer patients. *Journal of surgical oncology* 2013;108(1):9-13. doi: 10.1002/jso.23349 [published Online First: 2013/06/12]
 72. Rodel C, Graeven U, Fietkau R, et al. Oxaliplatin added to fluorouracil-based preoperative chemoradiotherapy and postoperative chemotherapy of locally advanced rectal cancer (the German CAO/ARO/AIO-04 study): final results of the multicentre, open-label, randomised, phase 3 trial. *The Lancet Oncology* 2015;16(8):979-89. doi: 10.1016/s1470-2045(15)00159-x [published Online First: 2015/07/21]
 73. Gray R, Barnwell J, McConkey C, et al. Adjuvant chemotherapy versus observation in patients with colorectal cancer: a randomised study. *Lancet (London, England)* 2007;370(9604):2020-9. doi: 10.1016/s0140-6736(07)61866-2 [published Online First: 2007/12/18]
 74. Cottet V, Bouvier V, Rollot F, et al. Incidence and patterns of late recurrences in rectal cancer patients. *Annals of surgical oncology* 2015;22(2):520-7. doi: 10.1245/s10434-014-3990-1 [published Online First: 2014/08/28]
 75. Nguyen DX, Bos PD, Massague J. Metastasis: from dissemination to organ-specific colonization. *Nature reviews Cancer* 2009;9(4):274-84. doi: 10.1038/nrc2622 [published Online First: 2009/03/25]

76. Rosenberg PS. Hazard Function Estimation Using B-Splines. *Biometrics* 1995;51(3):874-87. doi: 10.2307/2532989
77. de Boor C. A Practical Guide to Spline 1978.
78. Gilks WR, Wild P. Adaptive Rejection Sampling for Gibbs Sampling. *Journal of the Royal Statistical Society Series C (Applied Statistics)* 1992;41(2):337-48. doi: 10.2307/2347565
79. Rutter CM, Zaslavsky AM, Feuer EJ. Dynamic microsimulation models for health outcomes: a review. *Medical decision making : an international journal of the Society for Medical Decision Making* 2011;31(1):10-8. doi: 10.1177/0272989x10369005 [published Online First: 2010/05/21]
80. Hoeting JA, Madigan D, Raftery AE, et al. Bayesian model averaging: a tutorial (with comments by M. Clyde, David Draper and E. I. George, and a rejoinder by the authors. *Statist Sci* 1999;14(4):382-417. doi: 10.1214/ss/1009212519
81. Renfro LA, Sargent DJ. Findings from the Adjuvant Colon Cancer End Points (ACCENT) Collaborative Group: the power of pooled individual patient data from multiple clinical trials. *Chinese Clinical Oncology* 2016;5(6)
82. Goffe WL, Ferrier GD, Rogers J. Global optimization of statistical functions with simulated annealing. *Journal of Econometrics* 1994;60(1):65-99. doi: [https://doi.org/10.1016/0304-4076\(94\)90038-8](https://doi.org/10.1016/0304-4076(94)90038-8)
83. IntHout J, Ioannidis JPA, Rovers MM, et al. Plea for routinely presenting prediction intervals in meta-analysis. *BMJ open* 2016;6(7) doi: 10.1136/bmjopen-2015-010247
84. Ichikawa T, Saito K, Yoshioka N, et al. Detection and characterization of focal liver lesions: a Japanese phase III, multicenter comparison between gadoxetic acid disodium-enhanced magnetic resonance imaging and contrast-enhanced computed tomography predominantly in patients with hepatocellular carcinoma and chronic liver disease. *Investigative radiology* 2010;45(3):133-41. doi: 10.1097/RLI.0b013e3181caea5b [published Online First: 2010/01/26]
85. Niekel MC, Bipat S, Stoker J. Diagnostic imaging of colorectal liver metastases with CT, MR imaging, FDG PET, and/or FDG PET/CT: a meta-analysis of prospective studies including patients who have not previously undergone treatment. *Radiology* 2010;257(3):674-84. doi: 10.1148/radiol.10100729 [published Online First: 2010/09/11]
86. van Gestel YR, de Hingh IH, van Herk-Sukel MP, et al. Patterns of metachronous metastases after curative treatment of colorectal cancer. *Cancer epidemiology* 2014;38(4):448-54. doi: 10.1016/j.canep.2014.04.004 [published Online First: 2014/05/21]
87. Kuruppu D, Christophi C, Bertram JF, et al. Characterization of an animal model of hepatic metastasis. *Journal of gastroenterology and hepatology* 1996;11(1):26-32. [published Online First: 1996/01/01]
88. Finlay IG, Meek D, Brunton F, et al. Growth rate of hepatic metastases in colorectal carcinoma. *The British journal of surgery* 1988;75(7):641-4. [published Online First: 1988/07/01]
89. Benzekry S, Lamont C, Beheshti A, et al. Classical mathematical models for description and prediction of experimental tumor growth. *PLoS computational biology* 2014;10(8):e1003800. doi: 10.1371/journal.pcbi.1003800 [published Online First: 2014/08/29]
90. Iida T, Nomori H, Shiba M, et al. Prognostic factors after pulmonary metastasectomy for colorectal cancer and rationale for determining surgical indications: a retrospective

- analysis. *Annals of surgery* 2013;257(6):1059-64. doi: 10.1097/SLA.0b013e31826eda3b [published Online First: 2012/09/25]
91. Sadot E, Groot Koerkamp B, Leal JN, et al. Resection margin and survival in 2368 patients undergoing hepatic resection for metastatic colorectal cancer: surgical technique or biologic surrogate? *Annals of surgery* 2015;262(3):476-85; discussion 83-5. doi: 10.1097/sla.0000000000001427 [published Online First: 2015/08/11]
 92. Duineveld LA, van Asselt KM, Bemelman WA, et al. Symptomatic and Asymptomatic Colon Cancer Recurrence: A Multicenter Cohort Study. *Annals of family medicine* 2016;14(3):215-20. doi: 10.1370/afm.1919 [published Online First: 2016/05/18]
 93. Cleveland WS. Robust locally weighted regression and smoothing scatterplots. *Journal of the American statistical association* 1979;74(368):829-36.
 94. Flamen P, Hoekstra OS, Homans F, et al. Unexplained rising carcinoembryonic antigen (CEA) in the postoperative surveillance of colorectal cancer: the utility of positron emission tomography (PET). *European Journal of Cancer* 2001;37(7):862-69. doi: [https://doi.org/10.1016/S0959-8049\(01\)00049-1](https://doi.org/10.1016/S0959-8049(01)00049-1)
 95. McCall JL, Black RB, Rich CA, et al. The value of serum carcinoembryonic antigen in predicting recurrent disease following curative resection of colorectal cancer. *Diseases of the colon and rectum* 1994;37(9):875-81. [published Online First: 1994/09/01]
 96. Smith-Bindman R, Lipson J, Marcus R, et al. Radiation dose associated with common computed tomography examinations and the associated lifetime attributable risk of cancer. *Archives of Internal Medicine* 2009;169(22):2078-86. doi: 10.1001/archinternmed.2009.427
 97. Chew MH, Brown WE, Masya L, et al. Clinical, MRI, and PET-CT criteria used by surgeons to determine suitability for pelvic exenteration surgery for recurrent rectal cancers: a Delphi study. *Diseases of the colon and rectum* 2013;56(6):717-25. doi: 10.1097/DCR.0b013e3182812bec [published Online First: 2013/05/09]
 98. Young PE, Womeldorph CM, Johnson EK, et al. Early Detection of Colorectal Cancer Recurrence in Patients Undergoing Surgery with Curative Intent: Current Status and Challenges. *Journal of Cancer* 2014;5(4):262-71. doi: 10.7150/jca.7988
 99. Vogel WV, Wiering B, Corstens FHM, et al. Colorectal cancer: the role of PET/CT in recurrence: Wednesday 5 October 2005, 14:00–16:00. *Cancer Imaging* 2005;5(Spec No A):S143-S49. doi: 10.1102/1470-7330.2005.0034
 100. Valk PE, Abella-Columna E, Haseman MK, et al. Whole-body PET imaging with [18F]fluorodeoxyglucose in management of recurrent colorectal cancer. *Archives of surgery (Chicago, Ill : 1960)* 1999;134(5):503-11; discussion 11-3. [published Online First: 1999/05/14]
 101. Augestad KM, Norum J, Rose J, et al. A prospective analysis of false positive events in a National Colon Cancer Surveillance Program. *BMC health services research* 2014;14:137. doi: 10.1186/1472-6963-14-137 [published Online First: 2014/03/29]
 102. Litvak A, Cercek A, Segal N, et al. False-positive elevations of carcinoembryonic antigen in patients with a history of resected colorectal cancer. *Journal of the National Comprehensive Cancer Network : JNCCN* 2014;12(6):907-13. [published Online First: 2014/06/14]
 103. Stuckle CA, Haegele KF, Jendreck M, et al. [Improvements in detection of rectal cancer recurrence by multiplanar reconstruction]. *Der Radiologe* 2005;45(10):930-4, 36. doi: 10.1007/s00117-003-0950-3 [published Online First: 2005/10/28]

104. Maas M, Rutten IJ, Nelemans PJ, et al. What is the most accurate whole-body imaging modality for assessment of local and distant recurrent disease in colorectal cancer? A meta-analysis : imaging for recurrent colorectal cancer. *European journal of nuclear medicine and molecular imaging* 2011;38(8):1560-71. doi: 10.1007/s00259-011-1785-1 [published Online First: 2011/04/07]
105. Booth CM, Nanji S, Wei X, et al. Management and Outcome of Colorectal Cancer Liver Metastases in Elderly Patients: A Population-Based Study. *JAMA oncology* 2015;1(8):1111-9. doi: 10.1001/jamaoncol.2015.2943 [published Online First: 2015/09/12]
106. Landmann RG, Weiser MR. Surgical management of locally advanced and locally recurrent colon cancer. *Clinics in colon and rectal surgery* 2005;18(3):182-9. doi: 10.1055/s-2005-916279 [published Online First: 2005/08/01]
107. Poultsides GA, Schulick RD, Pawlik TM. Hepatic resection for colorectal metastases: the impact of surgical margin status on outcome. *HPB : the official journal of the International Hepato Pancreato Biliary Association* 2010;12(1):43-9. doi: 10.1111/j.1477-2574.2009.00121.x [published Online First: 2010/05/25]
108. Blackmon SH, Stephens EH, Correa AM, et al. Predictors of recurrent pulmonary metastases and survival after pulmonary metastasectomy for colorectal cancer. *The Annals of thoracic surgery* 2012;94(6):1802-9. doi: 10.1016/j.athoracsur.2012.07.014 [published Online First: 2012/10/16]
109. Hwang M, Jayakrishnan TT, Green DE, et al. Systematic review of outcomes of patients undergoing resection for colorectal liver metastases in the setting of extra hepatic disease. *European journal of cancer (Oxford, England : 1990)* 2014;50(10):1747-57. doi: 10.1016/j.ejca.2014.03.277 [published Online First: 2014/04/29]
110. John SK, Robinson SM, Rehman S, et al. Prognostic factors and survival after resection of colorectal liver metastasis in the era of preoperative chemotherapy: an 11-year single-centre study. *Digestive surgery* 2013;30(4-6):293-301. doi: 10.1159/000354310 [published Online First: 2013/08/24]
111. Tranchart H, Chirica M, Faron M, et al. Prognostic impact of positive surgical margins after resection of colorectal cancer liver metastases: reappraisal in the era of modern chemotherapy. *World journal of surgery* 2013;37(11):2647-54. doi: 10.1007/s00268-013-2186-3 [published Online First: 2013/08/29]
112. Laurent C, Adam JP, Denost Q, et al. Significance of R1 Resection for Advanced Colorectal Liver Metastases in the Era of Modern Effective Chemotherapy. *World journal of surgery* 2016;40(5):1191-9. doi: 10.1007/s00268-016-3404-6 [published Online First: 2016/01/14]
113. Lee KF, Wong J, Cheung YS, et al. Resection margin in laparoscopic hepatectomy: a comparative study between wedge resection and anatomic left lateral sectionectomy. *HPB : the official journal of the International Hepato Pancreato Biliary Association* 2010;12(9):649-53. doi: 10.1111/j.1477-2574.2010.00221.x [published Online First: 2010/10/22]
114. Leung U, Gonen M, Allen PJ, et al. Colorectal Cancer Liver Metastases and Concurrent Extrahepatic Disease Treated With Resection. *Annals of surgery* 2016 doi: 10.1097/sla.0000000000001624 [published Online First: 2016/01/15]
115. Truant S, Sequier C, Leteurtre E, et al. Tumour biology of colorectal liver metastasis is a more important factor in survival than surgical margin clearance in the era of modern

- chemotherapy regimens. *HPB : the official journal of the International Hepato Pancreato Biliary Association* 2015;17(2):176-84. doi: 10.1111/hpb.12316 [published Online First: 2014/07/22]
116. Eveno C, Karoui M, Gayat E, et al. Liver resection for colorectal liver metastases with peri-operative chemotherapy: oncological results of R1 resections. *HPB : the official journal of the International Hepato Pancreato Biliary Association* 2013;15(5):359-64. doi: 10.1111/j.1477-2574.2012.00581.x [published Online First: 2013/03/06]
117. Pandanaboyana S, White A, Pathak S, et al. Impact of margin status and neoadjuvant chemotherapy on survival, recurrence after liver resection for colorectal liver metastasis. *Annals of surgical oncology* 2015;22(1):173-9. doi: 10.1245/s10434-014-3953-6 [published Online First: 2014/08/03]
118. Ayez N, Lalmahomed ZS, Eggermont AM, et al. Outcome of microscopic incomplete resection (R1) of colorectal liver metastases in the era of neoadjuvant chemotherapy. *Annals of surgical oncology* 2012;19(5):1618-27. doi: 10.1245/s10434-011-2114-4 [published Online First: 2011/10/19]
119. National Center for Health Statistics. United States Life Tables, 2014. NVSR Volume 66, Number 4. 64pp. .
120. Cassidy J, Saltz L, Twelves C, et al. Efficacy of capecitabine versus 5-fluorouracil in colorectal and gastric cancers: a meta-analysis of individual data from 6171 patients. *Annals of oncology : official journal of the European Society for Medical Oncology* 2011;22(12):2604-9. doi: 10.1093/annonc/mdr031 [published Online First: 2011/03/19]
121. Sanoff HK, Sargent DJ, Campbell ME, et al. Five-year data and prognostic factor analysis of oxaliplatin and irinotecan combinations for advanced colorectal cancer: N9741. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 2008;26(35):5721-7. doi: 10.1200/jco.2008.17.7147 [published Online First: 2008/11/13]
122. Klein JP, Moeschberger ML. Survival analysis: techniques for censored and truncated data: Springer Science & Business Media 2005.
123. Pulitano C, Castillo F, Aldrighetti L, et al. What defines 'cure' after liver resection for colorectal metastases? Results after 10 years of follow-up. *HPB : the official journal of the International Hepato Pancreato Biliary Association* 2010;12(4):244-9. doi: 10.1111/j.1477-2574.2010.00155.x [published Online First: 2010/07/02]
124. Tomlinson JS, Jarnagin WR, DeMatteo RP, et al. Actual 10-year survival after resection of colorectal liver metastases defines cure. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 2007;25(29):4575-80. doi: 10.1200/jco.2007.11.0833 [published Online First: 2007/10/11]
125. Abbas S, Lam V, Hollands M. Ten-year survival after liver resection for colorectal metastases: systematic review and meta-analysis. *ISRN oncology* 2011;2011:763245. doi: 10.5402/2011/763245 [published Online First: 2011/11/18]
126. Kanas GP, Taylor A, Primrose JN, et al. Survival after liver resection in metastatic colorectal cancer: review and meta-analysis of prognostic factors. *Clinical Epidemiology* 2012;4:283-301. doi: 10.2147/CLEP.S34285
127. You YN, Skibber JM, Hu CY, et al. Impact of multimodal therapy in locally recurrent rectal cancer. *The British journal of surgery* 2016;103(6):753-62. doi: 10.1002/bjs.10079 [published Online First: 2016/03/05]

128. Liu W, Sun Y, Zhang L, et al. Negative surgical margin improved long-term survival of colorectal cancer liver metastases after hepatic resection: a systematic review and meta-analysis. *International journal of colorectal disease* 2015;30(10):1365-73. doi: 10.1007/s00384-015-2323-6 [published Online First: 2015/07/23]
129. Rabeneck L, Paszat LF, Li C. Risk factors for obstruction, perforation, or emergency admission at presentation in patients with colorectal cancer: a population-based study. *The American journal of gastroenterology* 2006;101(5):1098-103. doi: 10.1111/j.1572-0241.2006.00488.x [published Online First: 2006/04/01]
130. Khan MA, Hakeem AR, Scott N, et al. Significance of R1 resection margin in colon cancer resections in the modern era. *Colorectal disease : the official journal of the Association of Coloproctology of Great Britain and Ireland* 2015;17(11):943-53. doi: 10.1111/codi.12960 [published Online First: 2015/03/27]
131. Cai Y, Li Z, Gu X, et al. Prognostic factors associated with locally recurrent rectal cancer following primary surgery (Review). *Oncology Letters* 2014;7(1):10-16. doi: 10.3892/ol.2013.1640
132. Ikoma N, You YN, Bednarski BK, et al. Impact of Recurrence and Salvage Surgery on Survival After Multidisciplinary Treatment of Rectal Cancer. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 2017;35(23):2631-38. doi: 10.1200/jco.2016.72.1464 [published Online First: 2017/06/29]
133. Laubert T, Bader FG, Oevermann E, et al. Intensified surveillance after surgery for colorectal cancer significantly improves survival. *European journal of medical research* 2010;15(1):25-30. [published Online First: 2010/02/18]
134. Arriola E, Navarro M, Pares D, et al. Imaging techniques contribute to increased surgical rescue of relapse in the follow-up of colorectal cancer. *Diseases of the colon and rectum* 2006;49(4):478-84. doi: 10.1007/s10350-005-0280-9 [published Online First: 2006/02/02]
135. Upadhyay S, Dahal S, Bhatt VR, et al. Chemotherapy use in stage III colon cancer: a National Cancer Database analysis. *Therapeutic advances in medical oncology* 2015;7(5):244-51. doi: 10.1177/1758834015587867 [published Online First: 2015/09/04]
136. Becerra AZ, Probst CP, Tejani MA, et al. Opportunity lost: Adjuvant chemotherapy in patients with stage III colon cancer remains underused. *Surgery* 2015;158(3):692-9. doi: 10.1016/j.surg.2015.03.057 [published Online First: 2015/06/03]
137. Gelman A. Two simple examples for understanding posterior p-values whose distributions are far from uniform. *Electron J Statist* 2013;7:2595-602. doi: 10.1214/13-EJS854
138. Murthy VH, Krumholz HM, Gross CP. Participation in cancer clinical trials: Race-, sex-, and age-based disparities. *Jama* 2004;291(22):2720-26. doi: 10.1001/jama.291.22.2720
139. Treasure T, Fallowfield L, Lees B, et al. Pulmonary metastasectomy in colorectal cancer: the PulMiCC trial. *Thorax* 2012;67(2):185-7. doi: 10.1136/thoraxjnl-2011-200015 [published Online First: 2011/05/13]
140. Welch HG, Black WC. Overdiagnosis in Cancer. *JNCI: Journal of the National Cancer Institute* 2010;102(9):605-13. doi: 10.1093/jnci/djq099
141. Hansdotter Andersson P, Wille-Jørgensen P, Horváth-Puhó E, et al. The COLOFOL trial: study design and comparison of the study population with the source cancer population. *Clinical Epidemiology* 2016;8:15-21. doi: 10.2147/CLEP.S92661
142. Lepage C, Phelip JM, Cany L, et al. Effect of 5 years of imaging and CEA follow-up to detect recurrence of colorectal cancer: The FFCD PRODIGE 13 randomised phase III trial. *Dig*

- Liver Dis* 2015;47(7):529-31. doi: 10.1016/j.dld.2015.03.021 [published Online First: 2015/05/03]
143. Renehan AG, Egger M, Saunders MP, et al. Mechanisms of improved survival from intensive followup in colorectal cancer: a hypothesis. *British journal of cancer* 2005;92(3):430-3. doi: 10.1038/sj.bjc.6602369 [published Online First: 2005/02/03]
144. Expectancy or primary chemotherapy in patients with advanced asymptomatic colorectal cancer: a randomized trial. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 1992;10(6):904-11. doi: 10.1200/jco.1992.10.6.904 [published Online First: 1992/06/01]
145. Scheele J, Stangl R, Altendorf-Hofmann A. Hepatic metastases from colorectal carcinoma: impact of surgical resection on the natural history. *The British journal of surgery* 1990;77(11):1241-6. [published Online First: 1990/11/01]
146. Rodriguez-Moranta F, Salo J, Arcusa A, et al. Postoperative surveillance in patients with colorectal cancer who have undergone curative resection: a prospective, multicenter, randomized, controlled trial. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 2006;24(3):386-93. doi: 10.1200/jco.2005.02.0826 [published Online First: 2005/12/21]
147. Erenay FS, Alagoz O, Banerjee R, et al. Cost-effectiveness of alternative colonoscopy surveillance strategies to mitigate metachronous colorectal cancer incidence. *Cancer* 2016;122(16):2560-70. doi: 10.1002/cncr.30091 [published Online First: 2016/06/02]
148. Park SM, Kim SY, Earle CC, et al. What is the most cost-effective strategy to screen for second primary colorectal cancers in male cancer survivors in Korea? *World journal of gastroenterology* 2009;15(25):3153-60. [published Online First: 2009/07/04]
149. Renehan AG, O'Dwyer ST, Whynes DK. Cost effectiveness analysis of intensive versus conventional follow up after curative resection for colorectal cancer. *BMJ : British Medical Journal* 2004;328(7431):81-81.
150. Park KC, Schwimmer J, Shepherd JE, et al. Decision analysis for the cost-effective management of recurrent colorectal cancer. *Annals of surgery* 2001;233(3):310-9. [published Online First: 2001/02/27]
151. Rose J, Augestad KM, Kong CY, et al. A simulation model of colorectal cancer surveillance and recurrence. *BMC medical informatics and decision making* 2014;14:29. doi: 10.1186/1472-6947-14-29 [published Online First: 2014/04/09]
152. Lee-Ying RM, Kennecke HF, Nguyen L, et al. Cost-effectiveness of surveillance after curative resection (CR) of metastatic colorectal cancer (CRC). *Journal of Clinical Oncology* 2017;35(4_suppl):526-26. doi: 10.1200/JCO.2017.35.4_suppl.526
153. Gazelle GS, Hunink MGM, Kuntz KM, et al. Cost-Effectiveness of Hepatic Metastasectomy in Patients With Metastatic Colorectal Carcinoma: A State-Transition Monte Carlo Decision Analysis. *Annals of surgery* 2003;237(4):544-55. doi: 10.1097/01.SLA.0000059989.55280.33
154. Sanoff HK, Carpenter WR, Martin CF, et al. Comparative effectiveness of oxaliplatin vs non-oxaliplatin-containing adjuvant chemotherapy for stage III colon cancer. *J Natl Cancer Inst* 2012;104(3):211-27. doi: 10.1093/jnci/djr524 [published Online First: 2012/01/24]
155. Casadaban L, Rauscher G, Aklilu M, et al. Adjuvant chemotherapy is associated with improved survival in patients with stage II colon cancer. *Cancer* 2016;122(21):3277-87. doi: 10.1002/cncr.30181 [published Online First: 2016/10/21]

156. O'Connor ES, Greenblatt DY, LoConte NK, et al. Adjuvant chemotherapy for stage II colon cancer with poor prognostic features. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 2011;29(25):3381-8. doi: 10.1200/jco.2010.34.3426 [published Online First: 2011/07/27]
157. Razenberg LG, van Erning FN, Puijijt HF, et al. The impact of age on first-line systemic therapy in patients with metachronous metastases from colorectal cancer. *Journal of geriatric oncology* 2017;8(1):37-43. doi: 10.1016/j.jgo.2016.08.003 [published Online First: 2016/09/24]
158. Vargas GM, Sheffield KM, Parmar AD, et al. Trends in treatment and survival in older patients presenting with stage IV colorectal cancer. *Journal of gastrointestinal surgery : official journal of the Society for Surgery of the Alimentary Tract* 2014;18(2):369-77. doi: 10.1007/s11605-013-2406-z [published Online First: 2013/11/16]
159. Bradley CJ, Yabroff KR, Warren JL, et al. Trends in the Treatment of Metastatic Colon and Rectal Cancer in Elderly Patients. *Med Care* 2016;54(5):490-7. doi: 10.1097/mlr.0000000000000510 [published Online First: 2016/02/24]
160. National Comprehensive Cancer Network. Clinical practice guidelines in oncology: colon cancer. Version 2.2015 [Available from: http://www.nccn.org/professionals/physician_gls/pdf/colon.pdf accessed November 2015.
161. Cropper ML, Aydede SK, Portney PR. Discounting human lives. *American Journal of Agricultural Economics* 1991;73(5):1410-15.
162. Neumann PJ, Sanders GD, Ganiats TG, et al. Cost-Effectiveness in Health and Medicine: Oxford University Press 2016.
163. Agresti A. Categorical data analysis. 2nd ed. New York: New York : Wiley-Interscience 2002.
164. Lang K, Lines LM, Lee DW, et al. Lifetime and treatment-phase costs associated with colorectal cancer: evidence from SEER-Medicare data. *Clin Gastroenterol Hepatol* 2009;7(2):198-204. doi: 10.1016/j.cgh.2008.08.034 [published Online First: 2008/10/14]
165. Tom's Inflation Calculator. US Bureau of Labor Statistics.
166. Wright GE, Barlow WE, Green P, et al. Differences among the elderly in the treatment costs of colorectal cancer: how important is race? *Med Care* 2007;45(5):420-30. doi: 10.1097/01.mlr.0000257184.93944.80 [published Online First: 2007/04/21]
167. Brandi G, De Lorenzo S, Nannini M, et al. Adjuvant chemotherapy for resected colorectal cancer metastases: Literature review and meta-analysis. *World journal of gastroenterology* 2016;22(2):519-33. doi: 10.3748/wjg.v22.i2.519 [published Online First: 2016/01/27]
168. Warren JL, Yabroff KR, Meekins A, et al. Evaluation of Trends in the Cost of Initial Cancer Treatment. *JNCI: Journal of the National Cancer Institute* 2008;100(12):888-97. doi: 10.1093/jnci/djn175
169. Andres A, Mentha G, Adam R, et al. Surgical management of patients with colorectal cancer and simultaneous liver and lung metastases. *The British journal of surgery* 2015;102(6):691-9. doi: 10.1002/bjs.9783 [published Online First: 2015/03/20]
170. Battula N, Tsapralis D, Mayer D, et al. Repeat liver resection for recurrent colorectal metastases: a single-centre, 13-year experience. *HPB : the official journal of the International Hepato Pancreato Biliary Association* 2014;16(2):157-63. doi: 10.1111/hpb.12096 [published Online First: 2013/03/28]

171. Lam VW, Pang T, Laurence JM, et al. A systematic review of repeat hepatectomy for recurrent colorectal liver metastases. *Journal of gastrointestinal surgery : official journal of the Society for Surgery of the Alimentary Tract* 2013;17(7):1312-21. doi: 10.1007/s11605-013-2186-5 [published Online First: 2013/03/26]
172. Salah S, Watanabe K, Welter S, et al. Colorectal cancer pulmonary oligometastases: pooled analysis and construction of a clinical lung metastasectomy prognostic model. *Annals of oncology : official journal of the European Society for Medical Oncology* 2012;23(10):2649-55. doi: 10.1093/annonc/mds100 [published Online First: 2012/05/02]
173. Song X, Zhao Z, Barber B, et al. Characterizing medical care by disease phase in metastatic colorectal cancer. *J Oncol Pract* 2011;7(3 Suppl):25s-30s. doi: 10.1200/jop.2011.000304 [published Online First: 2011/09/03]
174. Lee SY, Oh SC. Advances of Targeted Therapy in Treatment of Unresectable Metastatic Colorectal Cancer. *Biomed Res Int* 2016;2016:7590245. doi: 10.1155/2016/7590245 [published Online First: 2016/04/30]
175. Frazier AL, Colditz GA, Fuchs CS, et al. Cost-effectiveness of screening for colorectal cancer in the general population. *Jama* 2000;284(15):1954-61. [published Online First: 2000/10/18]
176. Paulson EC, Veenstra CM, Vachani A, et al. Trends in surveillance for resected colorectal cancer, 2001-2009. *Cancer* 2015;121(19):3525-33. doi: 10.1002/cncr.29469 [published Online First: 2015/06/17]
177. Vargas GM, Sheffield KM, Parmar AD, et al. Physician follow-up and observation of guidelines in the post treatment surveillance of colorectal cancer. *Surgery* 2013;154(2):244-55. doi: 10.1016/j.surg.2013.04.013 [published Online First: 2013/07/31]
178. Sehdev A, Sherer EA, Hui SL, et al. Patterns of computed tomography surveillance in survivors of colorectal cancer at Veterans Health Administration facilities. *Cancer* 2017;123(12):2338-51. doi: 10.1002/cncr.30569 [published Online First: 2017/02/18]
179. Lansdorp-Vogelaar I, Knudsen A, Brenner H. Cost-effectiveness of colorectal cancer screening – an overview. *Best practice & research Clinical gastroenterology* 2010;24(4):439-49. doi: 10.1016/j.bpg.2010.04.004
180. Knudsen AB, Lansdorp-Vogelaar I, Rutter CM, et al. Cost-effectiveness of computed tomographic colonography screening for colorectal cancer in the medicare population. *J Natl Cancer Inst* 2010;102(16):1238-52. doi: 10.1093/jnci/djq242 [published Online First: 2010/07/29]
181. Claxton K. The irrelevance of inference: a decision-making approach to the stochastic evaluation of health care technologies. *Journal of health economics* 1999;18(3):341-64. [published Online First: 1999/10/28]
182. El-Shami K, Oeffinger KC, Erb NL, et al. American Cancer Society Colorectal Cancer Survivorship Care Guidelines. *CA: a cancer journal for clinicians* 2015;65(6):427-55. doi: 10.3322/caac.21286
183. Poston GJ, Tait D, O'Connell S, et al. Diagnosis and management of colorectal cancer: summary of NICE guidance. *BMJ (Clinical research ed)* 2011;343 doi: 10.1136/bmj.d6751
184. Kjeldsen BJ, Kronborg O, Fenger C, et al. A prospective randomized study of follow-up after radical surgery for colorectal cancer. *The British journal of surgery* 1997;84(5):666-9. [published Online First: 1997/05/01]

185. Wang T, Cui Y, Huang WS, et al. The role of postoperative colonoscopic surveillance after radical surgery for colorectal cancer: a prospective, randomized clinical study. *Gastrointestinal endoscopy* 2009;69(3 Pt 2):609-15. doi: 10.1016/j.gie.2008.05.017 [published Online First: 2009/01/13]
186. Verberne C, Zhan Z, Van Den Heuvel E, et al. Intensified follow-up in colorectal cancer patients using frequent Carcino-Embryonic Antigen (CEA) measurements and CEA-triggered imaging: Results of the randomized "CEAwatch" trial. *European Journal of Surgical Oncology* 2015;41(9):1188-96.
187. Verberne C, Zhan Z, van den Heuvel E, et al. Survival analysis of the CEAwatch multicentre clustered randomized trial. *British Journal of Surgery* 2017;104(8):1069-77.
188. Idrees JJ, Johnston FM, Canner JK, et al. Cost of Major Complications After Liver Resection in the United States: Are High-volume Centers Cost-effective? *Annals of surgery* 2017 doi: 10.1097/sla.0000000000002627 [published Online First: 2017/12/13]
189. Farjah F, Backhus LM, Varghese TK, et al. Ninety-day costs of video-assisted thoracic surgery versus open lobectomy for lung cancer. *The Annals of thoracic surgery* 2014;98(1):191-6. doi: 10.1016/j.athoracsur.2014.03.024 [published Online First: 2014/05/14]
190. Ash RB, Williams VL, Wagman LD, et al. Intraoperative radiotherapy for breast cancer: its perceived simplicity. *Oncology (Williston Park)* 2013;27(2):107-13. [published Online First: 2013/03/28]
191. Miller AR, Cantor SB, Peoples GE, et al. Quality of life and cost effectiveness analysis of therapy for locally recurrent rectal cancer. *Diseases of the colon and rectum* 2000;43(12):1695-701; discussion 701-3. [published Online First: 2001/01/13]
192. Ayvaci MU, Shi J, Alagoz O, et al. Cost-effectiveness of adjuvant FOLFOX and 5FU/LV chemotherapy for patients with stage II colon cancer. *Medical decision making : an international journal of the Society for Medical Decision Making* 2013;33(4):521-32. doi: 10.1177/0272989x12470755 [published Online First: 2013/01/15]
193. Zaydfudim VM, McMurry TL, Harrigan AM, et al. Improving treatment and survival: a population-based study of current outcomes after a hepatic resection in patients with metastatic colorectal cancer. *HPB : the official journal of the International Hepato Pancreato Biliary Association* 2015;17(11):1019-24. doi: 10.1111/hpb.12488 [published Online First: 2015/09/12]
194. Cho JH, Kim S, Namgung M, et al. The prognostic importance of the number of metastases in pulmonary metastasectomy of colorectal cancer. *World J Surg Oncol* 2015;13:222. doi: 10.1186/s12957-015-0621-7 [published Online First: 2015/07/25]
195. Karim S, Nanji S, Brennan K, et al. Chemotherapy for resected colorectal cancer pulmonary metastases: Utilization and outcomes in routine clinical practice. *European journal of surgical oncology : the journal of the European Society of Surgical Oncology and the British Association of Surgical Oncology* 2017;43(8):1481-87. doi: 10.1016/j.ejso.2017.05.003 [published Online First: 2017/06/22]
196. Brunelli A, Drosos P, Dinesh P, et al. The Severity of Complications Is Associated With Postoperative Costs After Lung Resection. *The Annals of thoracic surgery* 2017;103(5):1641-46. doi: 10.1016/j.athoracsur.2016.10.061 [published Online First: 2017/02/13]

197. Welter S, Jacobs J, Krbek T, et al. Long-term survival after repeated resection of pulmonary metastases from colorectal cancer. *The Annals of thoracic surgery* 2007;84(1):203-10. doi: 10.1016/j.athoracsur.2007.03.028 [published Online First: 2007/06/26]
198. Hamady ZZR, Kotru A, Nishio H, et al. Current techniques and results of liver resection for colorectal liver metastases. *British Medical Bulletin* 2004;70(1):87-104. doi: 10.1093/bmb/ldh026
199. Krishnamurthy A, Kankesan J, Wei X, et al. Chemotherapy delivery for resected colorectal cancer liver metastases: Management and outcomes in routine clinical practice. *European Journal of Surgical Oncology (EJSO)* 2017;43(2):364-71. doi: <https://doi.org/10.1016/j.ejso.2016.08.022>
200. Bowne WB, Lee B, Wong WD, et al. Operative salvage for locoregional recurrent colon cancer after curative resection: an analysis of 100 cases. *Diseases of the colon and rectum* 2005;48(5):897-909. doi: 10.1007/s10350-004-0881-8 [published Online First: 2005/03/24]
201. Park JS, Kim HK, Choi YS, et al. Outcomes after repeated resection for recurrent pulmonary metastases from colorectal cancer. *Annals of oncology : official journal of the European Society for Medical Oncology* 2010;21(6):1285-9. doi: 10.1093/annonc/mdp475 [published Online First: 2009/10/29]

Appendix A

Table A.1: Most Recent Professional Society Guidelines for Follow-up

Organization (Reference)	Stages Covered	Endoscopy (Intraluminal)	CEA	Clinical Visit (History & Physical)	Imaging
ASCO ¹³	II, III	Colonoscopy at 1 year after surgery (before 1 year if no clearing colonoscopy) and then every 5 years (depending upon results).	Every 3-6 months for 5 years (depending on risk)	Every 3-6 months for 5 years (depending on risk)	Annual CT of Abdomen and Chest for 3 years (6 months if high risk). For high-risk rectal, also Pelvic CT
ASCRS ¹⁴	I (high-risk), II, III, IV (isolated metastases treated for cure)	Colonoscopy at 1 year (3-6 months if no clearing colonoscopy) and then in 1 (3) more years if adenomas (normal). For Rectal cancer, proctoscopy every 6-12 months for 3-5 years	Every 3-6 months for 2 years, and every 6 months until 5 years	Every 3-6 months for 2 years, and every 6 months until 5 years	Annual CT of abdomen, chest, and pelvis for 5 years for colon and rectal (6 months if high risk)
NCCN ¹⁶⁰	I-III	Colonoscopy at 1 year and then 3 years later and then every 5 years if clear	Every 3-6 months for 2 years, and every 6 months until 5 years	Every 3-6 months for 2 years, and every 6 months until 5 years	Annual CT of abdomen, chest, and pelvis for colon and rectal cancer for up to 5 years (depending on risk)
ACS ¹⁸²	I-III	Colonoscopy at year 1 and then repeat in 3 years (1 year) if clear (advanced adenoma). If clear at year 4, then repeat every 5 years.	Every 3-6 months for 2 years, and every 6 months until 5 years	Every 3-6 months for 2 years, and every 6 months until 5 years	Annual CT of abdomen, chest, and pelvis for colon and rectal cancer for 5 years (only high risk stage I-II)

NICE (UK) ¹⁸³	I-III, IV (treated with potentially curative intent)	Colonoscopy at 1 year. If normal, another colonoscopy in 5 years. If abnormal, interval depends upon finding.	At least every 6 months for first 3 years.	Regular follow-up, beginning 4-6 weeks after operation.	CT of abdomen, chest, and pelvis at least 2 times in first 3 years for rectal and colon cancer.
ESMO ¹⁶	Stage I-III, Stage IV if treated with curative intent. Imaging for stage II-IV generally	At diagnosis and then every 5 years if clear. For rectal cancer treated with local excision, sigmoidoscopy every 3-6 months for 3 years and then 6-12 months until 5 years.	Every 3 months for 2 years, and every 6 months until 5 years	Every 3 months for 2 years, and every 6 months until 5 years	CT of abdomen and chest every 6 months (stage IV) or 12 months (stage II-III) for high risk patients (maybe stopping at 3 years). Might replace CT of abdomen with contrast-enhanced ultrasound every 3-6 months.

ASCO: American Society of Clinical Oncology. ASCRS: American Society of Colon and Rectal Surgeons.

NCCN: National Cancer Care Network. ACS: American Cancer Society. NICE: National Institute for Health and Care Excellence (UK). ESMO: European Society for Medical Oncology.

Table A.2: Details of Randomized Control Trials

Lead Author (Year of Publication) & Source	Sample Size	Enrollment Period	Country	Location Distribution	Stage Distribution	Overall Recurrence
Makela (1995) ³¹	Trt N = 52 Ctrl N = 54	1988-1990	Finland	Trt Rectal = 30% Ctrl Rectal = 28%	Dukes A = 26% Dukes B = 45% Dukes C = 28%	41%
Ohlsson (1995) ³²	Trt N = 53 Ctrl N = 54	1983-1986	Sweden	Trt Rectal = 36% Ctrl Rectal = 32%	Dukes A = 18% Dukse B = 44% Dukes C = 38%	33%
Kjeldson (1997) ¹⁸⁴	Trt N = 290 Ctrl N = 307	1983-1994	Denmark	Trt Rectal = 46% Ctrl Rectal = 49%	Dukes A = 23% Dukes B = 49% Dukes C = 28%	24%
Pietra (1998) ³³	Trt N = 104 Ctrl N = 103	1987-1990	Italy	Trt Rectal = 30% Ctrl rectal = 36%	Dukes B = 59% Dukes C = 41%	40%
Schoemaker (1998) ³⁴	Trt N = 167 Ctrl N = 158	1984-1990	US	Trt Rectal = 28% Ctrl Rectal = 26%	Dukes A = 22% Dukes B = 47% Dukes C = 31%	37%
Secco (2002) ⁴¹	Trt N = 192 (108 high risk, 84 low risk) Ctrl N = 145 (84 high risk, 61 low risk)	1988-1996	Italy	Trt Rectal = 100% Ctrl Rectal = 100%	Dukes B = 40% Dukes C = 60%	55%
Rodriquez-Moranta (2006) ¹⁴⁶	Trt N = 127 Ctrl N = 132	1997-2001	Spain	Trt Rectal = 23% Ctrl Rectal = 27%	Stage II = 61% Stage III = 39%	27%
Sobhani (2008) ²⁸	Trt N = 65 Ctrl N = 65	2001-2004	France	Trt Rectal = 44% Ctrl Rectal = 41%	Unclear	35% (Only 15 months follow-up)
Wang (2009) ¹⁸⁵	Trt N = 165 Ctrl N = 161	1995-2001	China	Trt Rectal = 47% Ctrl Rectal = 48%	Dukes A = 31% Dukes B = 41% Dukes C = 29%	Unclear
Primrose (2014) ^{45 46} "FACS Trial"	CT & CEA N = 302 CT Only N = 299 CEA Only N =	2003-2009	UK	CEA & CT Rectal = 32% CT-Only Rectal = 34% CEA-Only	Dukes A = 22% Dukes B = 47%	17%

	300 Min N = 301			Rectal = 28% Minimal Arm = 30%	Dukes C = 31%	
Rosati (2016) ³⁵ "GILDA Trial"	Trt N = 615 Ctrl N = 613	1998-2006	Italy	Trt Rectal = 24% Ctrl Rectal = 24%	Dukes B = 50% Dukes C = 50%	23%
Treasure (2014) ²⁴ "CEA Second Look Trial"	Trt N = 108 Ctrl N = 108	1982-1993	UK	Unclear	Dukes A = 5% Dukes B = 46% Dukes C = 49%	Unclear
Verbone (2015) ^{186 187} "CEAwatch Trial"	Used Stepped- Wedge Cluster Design Hospital N = 11 Patient N = 3,223	2010-2012	Netherlands	Overall = 37%	Dukes A = 67% Dukes B = 33%	8%

Table A.2 gives details of the 13 randomized control trials that have compared more intensive to less intensive follow-up regimens. Trt = Treatment-arm; Ctrl = Control-arm

Table A.3: Surveillance Schedules of Randomized Control Trials

Lead Author (Year of Publication) & Source	Imaging	CEA Assays (Medical Exam and Physical History Implied)	Endoscopy
Makela (1995) ³¹	<p>Trt: <i>Chest X-ray</i>: every 3 months for 2 years, every 6 months for 3 years <i>Ultrasound of Liver</i>: Every 6 months for 5 years <i>Abdominal CT</i>: Annually for 5 years (pelvic for rectal likely too, not clear) Ctrl: <i>Chest X-ray</i>: every 3 months for 2 years, every 6 months for 3 years</p>	<p>Trt: every 3 months for 2 years, every 6 months for 3 years Ctrl: every 3 months for 2 years, every 6 months for 3 years</p>	<p>Trt: <i>Colonoscopy</i>: Annually year 1-5 <i>Flexible Sigmoidoscopy</i> (Rectal and Sigmoid Only): every 3 months for 5 years Ctrl: <i>Barium Enema</i>: Annually year 1-5 <i>Rigid Sigmoidoscopy</i> (Rectal and Sigmoid Only): every 3 months for 2 years, every 6 months for 3 years</p>
Ohlsson (1995) ³²	<p>Trt: <i>Chest X-ray</i>: every 3 months for 2 years, every 6 months for 2 years, and at 5 years <i>CT of pelvis</i>: (For rectal patients only): at 3, 6, 12, 18, 24 months Ctrl: ---</p>	<p>Trt: every 3 months for 2 years, every 6 months for 2 years, and at 5 years Ctrl: ---</p>	<p>Trt: <i>Colonoscopy or flexible sigmoidoscopy</i>: at 3, 9, 15, 21, 30, 42, 60 months Ctrl: Recommendation to get occult blood testing every 3 months for 2 years, and then once a year for 3 years</p>
Kjeldson (1997) ¹⁸⁴	<p>Trt: <i>Chest X-ray</i>: every 6 months for 3 years, every 12 months for 2 years, 10 years, 12.5 years, 15 years Ctrl: <i>Chest X-ray</i>: 5 years, 10 years, 15, years</p>	<p>Trt: --- Ctrl: ---</p>	<p>Trt: <i>Colonoscopy</i>: every 6 months for 3 years, every 12 months for 2 years, 10 years, 12.5 years, 15 years Ctrl: <i>Colonoscopy</i>: 5 years, 10 years, 15, years</p>
Pietra (1998) ³³	<p>Trt: <i>Abomdinal US</i>: every 3 months for 2 years, every 6 months for 3 years <i>CT</i> (unclear what type): annually for 5 years <i>Chest X-ray</i>: annually for 5 years Ctrl: <i>Abdominal US</i>: every 6 months for 1 year, annually for 4 years</p>	<p>Trt: every 3 months for 2 years, every 6 months for 3 years Ctrl: every 6 months for 1 year, then annually for 4 years</p>	<p>Trt: <i>Colonoscopy</i>: annually for 5 years Ctrl: <i>Colonoscopy</i>: annually for 5 years</p>

	<i>Chest X-ray</i> : annually for 5 years		
Schoemaker (1998) ³⁴	Trt: <i>Abdominal CT</i> : annually for 5 years <i>Chest X-ray</i> : annually for 5 years Ctrl: ---	Trt: every 3 months for 2 years, every 6 months for 3 years Ctrl: every 3 months for 2 years, every 6 months for 3 years	Trt: <i>Colonoscopy</i> : annually for 5 years Ctrl: ---
Secco (2002) ⁴¹ High Risk Subset	Trt: <i>Abdominal/Pelvic US</i> : every 6 months for 3 years, annually for 2 years <i>Chest X-ray</i> : annually for 5 years Ctrl: ---	Trt: every 3 months for 2 years, every 4 months for 3 rd year, every 6 months for 4 th /5 th year Ctrl: ---	Trt: <i>Rigid Sigmoidoscopy</i> : annually for 5 years Ctrl: ---
Secco (2002) ⁴¹ Low Risk Subset	Trt: <i>Abdominal/Pelvic US</i> : every 6 months for 2 years, annually for 3 years <i>Chest X-ray</i> : annually for 5 years Ctrl: ---	Trt: every 6 months for 2 years, annually for 3 years Ctrl: ---	Trt: <i>Rigid Sigmoidoscopy</i> : years 1, 2, 4 Ctrl: ---
Rodriguez-Moranta (2006) ¹⁴⁶	Trt: <i>Abdominal US</i> : (colon only) every 6 months for 2 years, annually for 3 years <i>Abdominal/Pelvic CT</i> : (rectal only) every 6 months for 2 years, annually for 3 years <i>Chest X-ray</i> : annually for 5 years Ctrl: ---	Trt: every 3 months for 2 years, every 6 months for 3 years Ctrl: every 3 months for 2 years, every 6 months for 3 years	Trt: <i>Colonoscopy</i> : annually for 5 years Ctrl: <i>Colonoscopy</i> : years 1 and 3
Sobhani (2008) ²⁸ (trial only involved only 15 months of follow-up: from 9 to 24 months post surgery)	Trt: <i>Abdominal US</i> : at 12, 18, 21, 24 months <i>Abdominal CT</i> : at 9, 15 months <i>FDG-PET</i> : at 9, 15 months <i>Chest X-ray</i> : at 9, 15, 21 months Ctrl: <i>Abdominal US</i> : at 12, 18, 21, 24 months <i>Abdominal CT</i> : at 9, 15 months <i>Chest X-ray</i> : at 9, 15, 21 months	Trt: At 9, 12, 15, 18, 21, 24 months Ctrl: At 9, 12, 15, 18, 21, 24 months	Trt: --- Ctrl: ---
Wang (2009) ¹⁸⁵	Trt: <i>Abdominal US or CT</i> : every 3 months for 1 year, every 6 months for 2 years, annually for 2 years	Trt: every 3 months for 1 year, every 6 months for 2 years, annually for 2 years Ctrl:	Trt: <i>Colonoscopy</i> : every 3 months for 1 year, every 6 months for 2 years, annually for 2 years

	<p><i>Chest X-ray</i>: every 3 months for 1 year, every 6 months for 2 years, annually for 2 years</p> <p>Ctrl:</p> <p><i>Abdominal US or CT</i>: every 3 months for 1 year, every 6 months for 2 years, annually for 2 years</p> <p><i>Chest X-ray</i>: every 3 months for 1 year, every 6 months for 2 years, annually for 2 years</p>	<p>every 3 months for 1 year, every 6 months for 2 years, annually for 2 years</p>	<p>Ctrl:</p> <p><i>Colonoscopy</i>: at 6, 30, 60 months</p>
<p>Primrose (2014)^{45 46} "FACS Trial"</p>	<p>CT & CEA:</p> <p><i>CT of Chest/Abdomen/Pelvis</i>: every 6 months for 2 years, annually for 3 years</p> <p>CT Only:</p> <p><i>CT of Chest/Abdomen/Pelvis</i>: every 6 months for 2 years, annually for 3 years</p> <p>CEA Only:</p> <p><i>CT of Chest/Abdomen/Pelvis</i>: 1 scan at 12-18 months</p> <p>Minimal:</p> <p><i>CT of Chest/Abdomen/Pelvis</i>: 1 scan at 12-18 months</p>	<p>CT & CEA:</p> <p>every 3 months for 2 years, every 6 months for 3 years</p> <p>CT Only:</p> <p>---</p> <p>CEA Only:</p> <p>every 3 months for 2 years, every 6 months for 3 years</p> <p>Minimal:</p> <p>---</p>	<p>CT & CEA:</p> <p><i>Colonoscopy</i>: 2 years and 5 years</p> <p>CT Only:</p> <p><i>Colonoscopy</i>: 2 years and 5 years</p> <p>CEA Only:</p> <p><i>Colonoscopy</i>: 2 years and 5 years</p> <p>Minimal:</p> <p><i>Colonoscopy</i>: at 5 years</p>
<p>Rosati (2016)³⁵ "GILDA Trial"</p>	<p>Trt:</p> <p><i>Abdominal US</i>: at 4, 8, 12, 16, 24, 36, 48, 60 months</p> <p><i>Abdominal/Pelvic CT</i>: (rectal only) at 4, 12, 24, 48 months</p> <p><i>Chest X-ray</i>: annually for 5 years</p> <p>Ctrl:</p> <p><i>Abdominal US</i>: (colon) at 4, 16 months</p> <p><i>Abdominal US</i>: (rectal) at 8, 16 months</p> <p><i>Chest X-ray</i>: (rectal only) at 1 year</p>	<p>Trt:</p> <p>every 4 months for 2 years, every 6 months for 2 years, at 5 years</p> <p>Ctrl:</p> <p>every 4 months for 2 years, every 6 months for 2 years, at 5 years</p>	<p>Trt:</p> <p><i>Colonoscopy</i>: annually for 5 years</p> <p><i>Proctoscopy</i>: (rectal only) at 4 and 8 months</p> <p>Ctrl:</p> <p><i>Colonoscopy</i>: years 1 and 4</p> <p><i>Proctoscopy</i>: (rectal only) at 4 months</p>
<p>Treasure (2014)²⁴ "CEA Second Look Trial"</p>	<p>Trt:</p> <p>---</p> <p>Ctrl:</p> <p>---</p>	<p>All randomized patients had elevated CEA-levels. They were then randomized to second-look surgery or conservative treatment</p>	<p>Trt:</p> <p>---</p> <p>Ctrl:</p> <p>---</p>
<p>Verbone (2015)^{186 187} "CEAwatch Trial"</p>	<p>Trt:</p> <p><i>Abdominal CT</i>: annually for 3 years</p> <p>Ctrl:</p> <p><i>Abdominal US</i>: every 6 months for 3 years, annually for 2 years</p> <p><i>Chest X-ray</i>: every 6 months for 3 years, annually for 2 years</p>	<p>Trt:</p> <p>every 2 months for 3 years, every 3 months for 2 years</p> <p>Ctrl:</p> <p>every 3-6 months for 3 years, annually for 2 years</p>	<p>Trt:</p> <p>---</p> <p>Ctrl:</p> <p>---</p>

US = Ultrasound; Trt = Treatment Arm; Ctrl = Control Arm;

Appendix B

Figure B.1: Forest Plot of the Log Odds Ratio of the Chances of Asymptomatic Detection

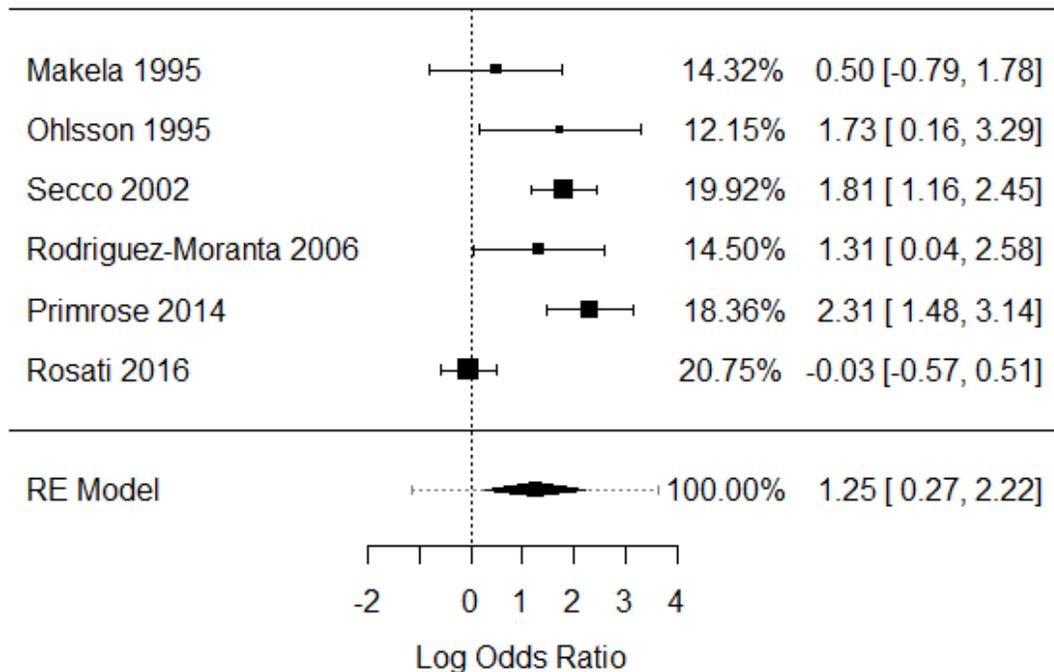


Figure B.1 depicts the forest plot of the log odds ratio (OR) of the chances of asymptomatic (surveillance-based) detection of disease comparing intensive to control arms. A log OR of > 0 signifies the intensive arm had a greater proportion of surveillance-detected recurrences. The first column on the right gives the weights for each trial in the random-effect model. The dashed line that runs horizontally through the pooled-effect diamond shows the limits of the prediction interval.

Of the 13 potentially relevant studies, two (Kjeldson 1997 & Wang 2009) were excluded due to a focus on endoscopic follow-up. Sobhani (2008) was excluded due to limited follow-up time (15 months only). The CEASL trial was excluded for questionable

relevance (unclear if CEA lead to CT scans to confirm recurrence), and the CEAwatch trial was excluded because it did not use a fully experimental design. Finally, two potentially relevant trials (Pietra 1998 and Schoemaker 1998) were excluded due to inadequate information.

Figure B.2: Forest Plot of the Probability of Asymptomatic Detection in the Intervention Arm

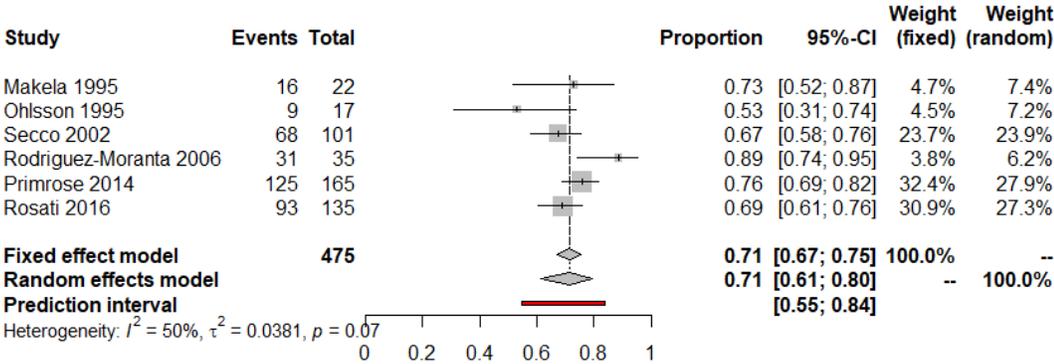


Figure B.2 depicts a forest plot for the probability that a patient with recurrence is detected asymptotically in the intensive arm. It uses the same includes/excludes as Figure B.1.

Figure B.3: Forest Plot of the Log Odds Ratio of the Chances of R0 Curative Resection among Patients with a Recurrence

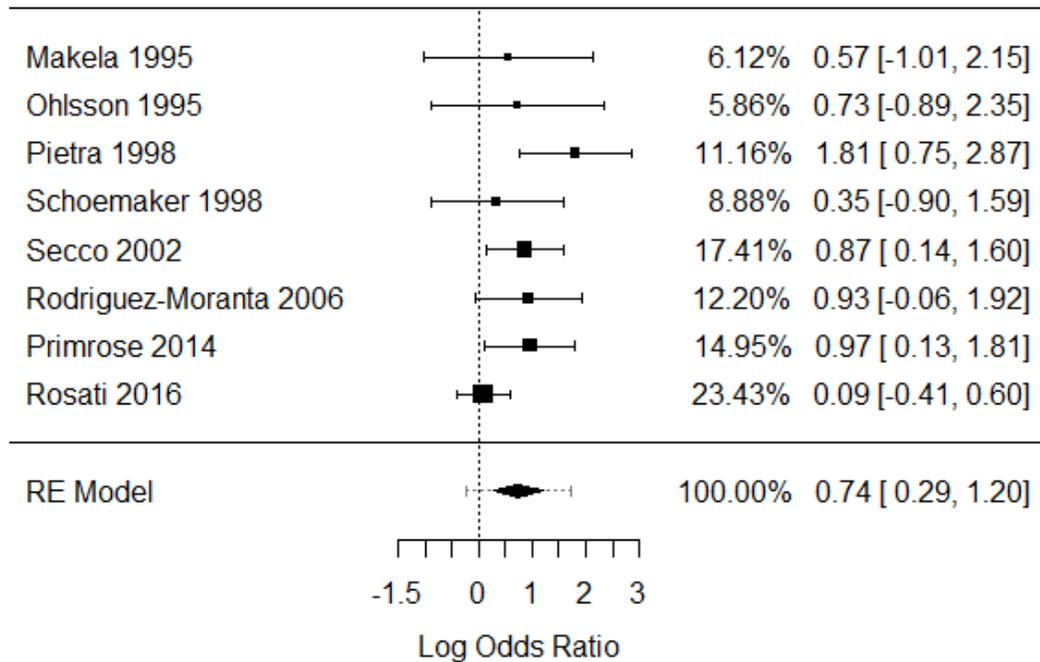


Figure B.3 depicts the forest plot of the log odds ratio (OR) of the chances of R0 curative resection among patients who present with a recurrence within 5 years comparing intensive to control arms. A log OR of > 0 signifies the intensive arm had a greater proportion of R0 curative resections. Recall an R0 resection means macroscopically and microscopically clear margins. The first column on the right gives the weights for each trial in the random-effect model. The dashed line that runs horizontally through the pooled-effect diamond shows the limits of the prediction interval.

Of the 13 potentially relevant studies, 2 (Kjeldson 1997 & Wang 2009) were excluded due to a focus on endoscopic follow-up. Sobhani (2008) was excluded due to limited follow-up. The CEASL trial was excluded for questionable relevance (unclear if CEA lead to CT scans to confirm recurrence), and the CEAwatch trial was excluded because it did not technically use an experimental design.

Figure B.4: Forest Plot of the Probability of R0 Curative Resection in the Intervention Arm

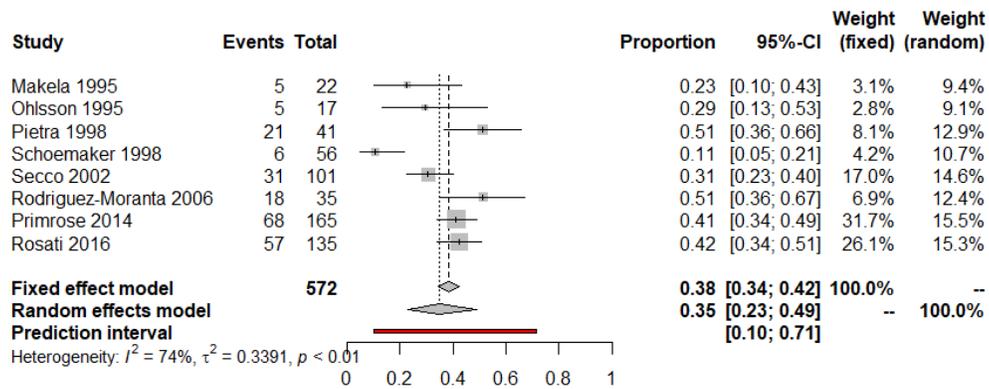


Figure B.4 depicts a forest plot for the probability that a patient with recurrence undergoes R0 curative resection in the intensive arm. It uses the same includes/excludes as Figure B.3.

Figure B.5: Stage III Colon Cancer No Adjuvant Therapy

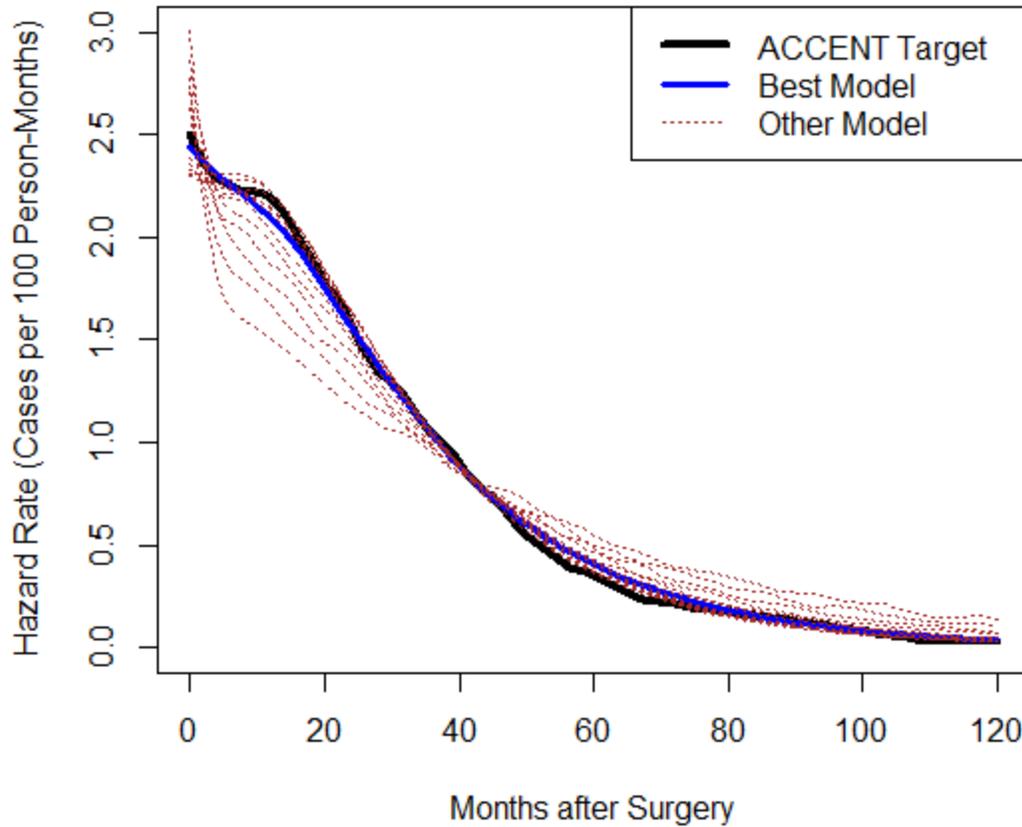


Figure B.5 depicts the calibration results for stage III colon cancer patients treated with surgery alone. The ‘Best Model’ line is the model-output (smoothed hazard function) using the highest-probability parameter sets (for the time-to-detectable-disease and time-to-clinical-disease processes). The brown dashed lines represent the model-output with the other included parameter sets. The y-axis presents the hazard rate and has been scaled to cases per 100-person months.

The data for this target was assumed to have been generated in the absence of any routine follow-up.

Figure B.6: Stage III Colon Cancer 5-FU + LV

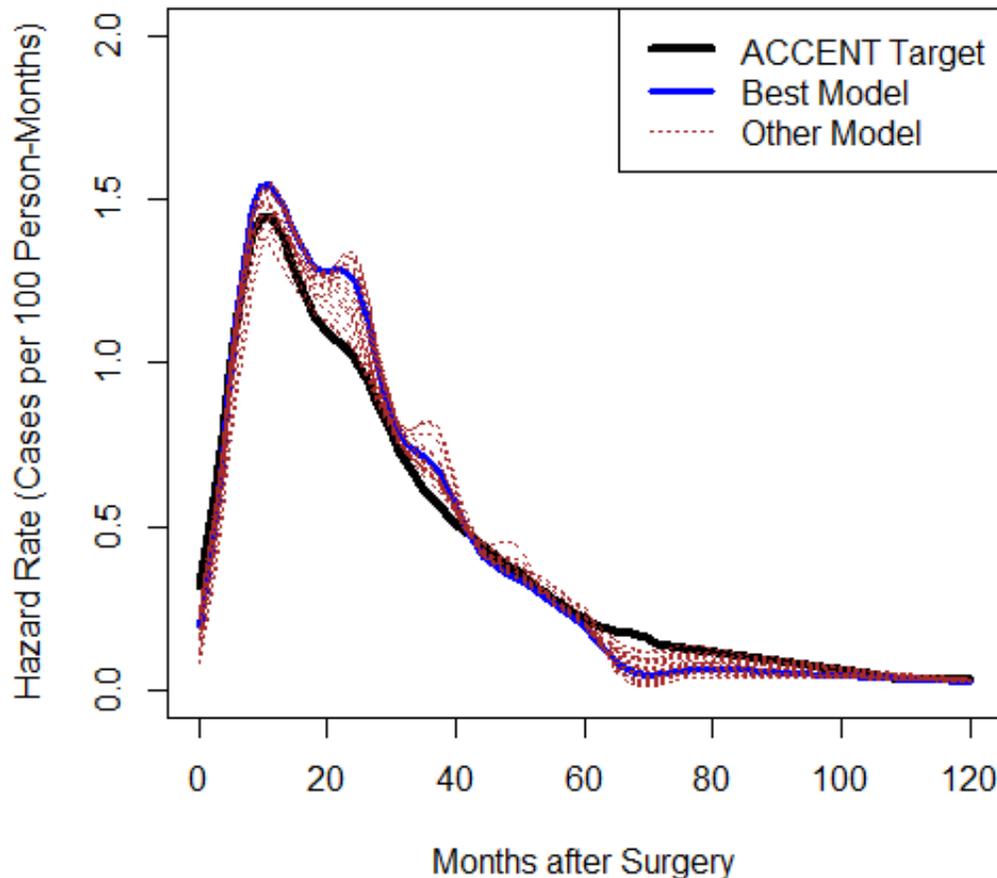


Figure B.6 depicts the calibration results for stage III colon cancer patients treated with adjuvant 5-FU/LV. The ‘Best Model’ line is the model-output (smoothed hazard function) using the highest-probability parameter sets (for the time-to-detectable-disease and time-to-clinical-disease processes). The brown dashed lines represent the model-output with the other included parameter sets. The y-axis presents the hazard rate and has been scaled to cases per 100-person months.

The data for this target was assumed to have been generated with under the following follow-up schedule: CEA every 3 months for 2 years, then every 6 months for 3 years. CT every year for 5 years.

Figure B.7: Stage III Colon Cancer FOLFOX

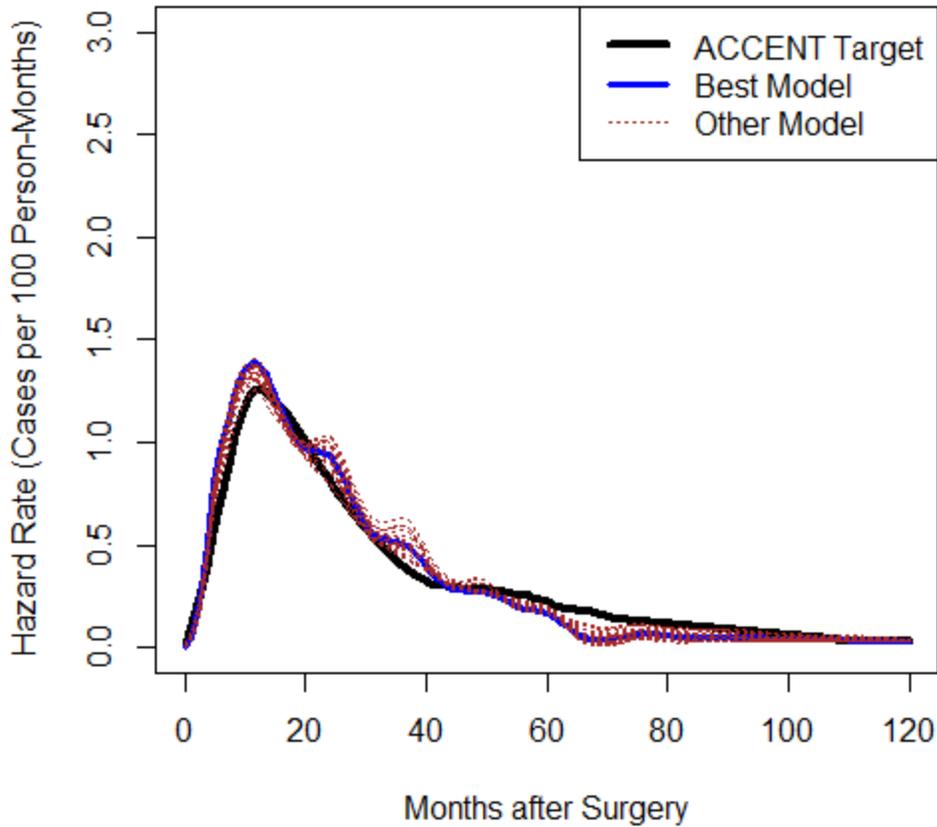


Figure B.7 depicts the calibration results for stage III colon cancer patients treated with adjuvant FOLOFX. The ‘Best Model’ line is the model-output (smoothed hazard function) using the highest-probability parameter sets (for the time-to-detectable-disease and time-to-clinical-disease processes). The brown dashed lines represent the model-output with the other included parameter sets. The y-axis presents the hazard rate and has been scaled to cases per 100-person months.

The data for this target was assumed to have been generated with under the following follow-up schedule: CEA every 3 months for 2 years, then every 6 months for 3 years. CT every year for 5 years.

Figure B.8: (Old) Stage II Colon Cancer No Adjuvant

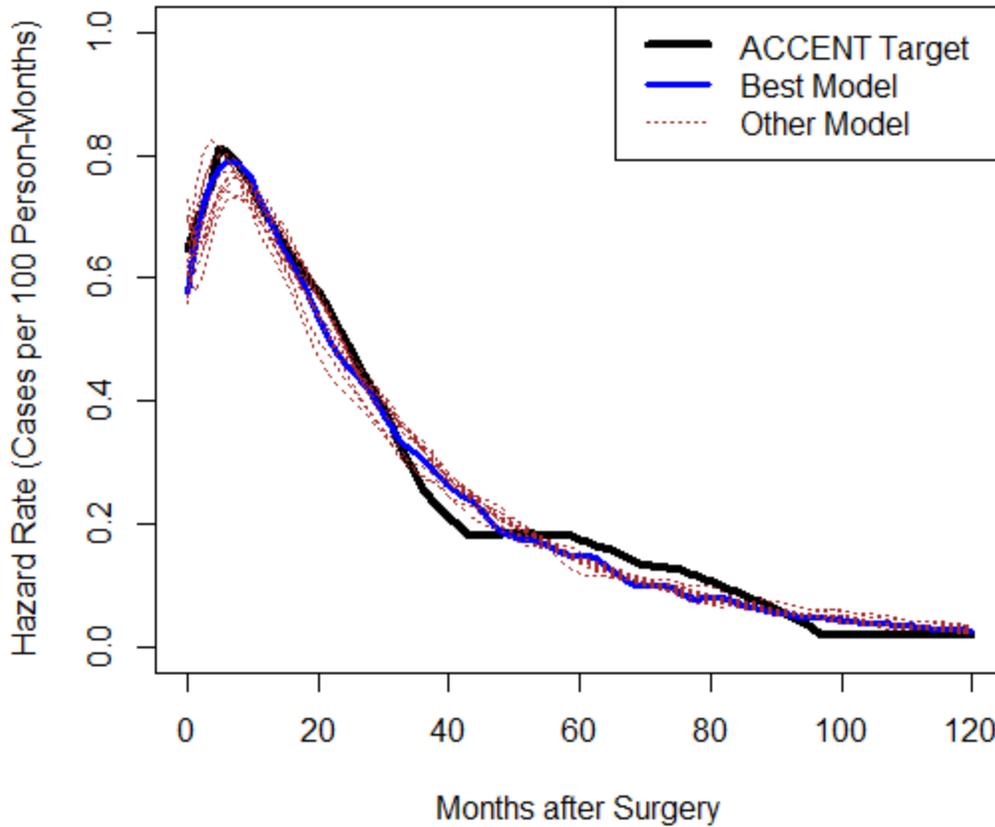


Figure B.8 depicts the calibration results for stage II colon cancer patients treated with surgery alone in the old era. The ‘Best Model’ line is the model-output (smoothed hazard function) using the highest-probability parameter sets (for the time-to-detectable-disease and time-to-clinical-disease processes). The brown dashed lines represent the model-output with the other included parameter sets. The y-axis presents the hazard rate and has been scaled to cases per 100-person months.

The data for this target was assumed to have been generated in the absence of any routine follow-up.

Figure B.9: (New) Stage II Colon Cancer No Adjuvant

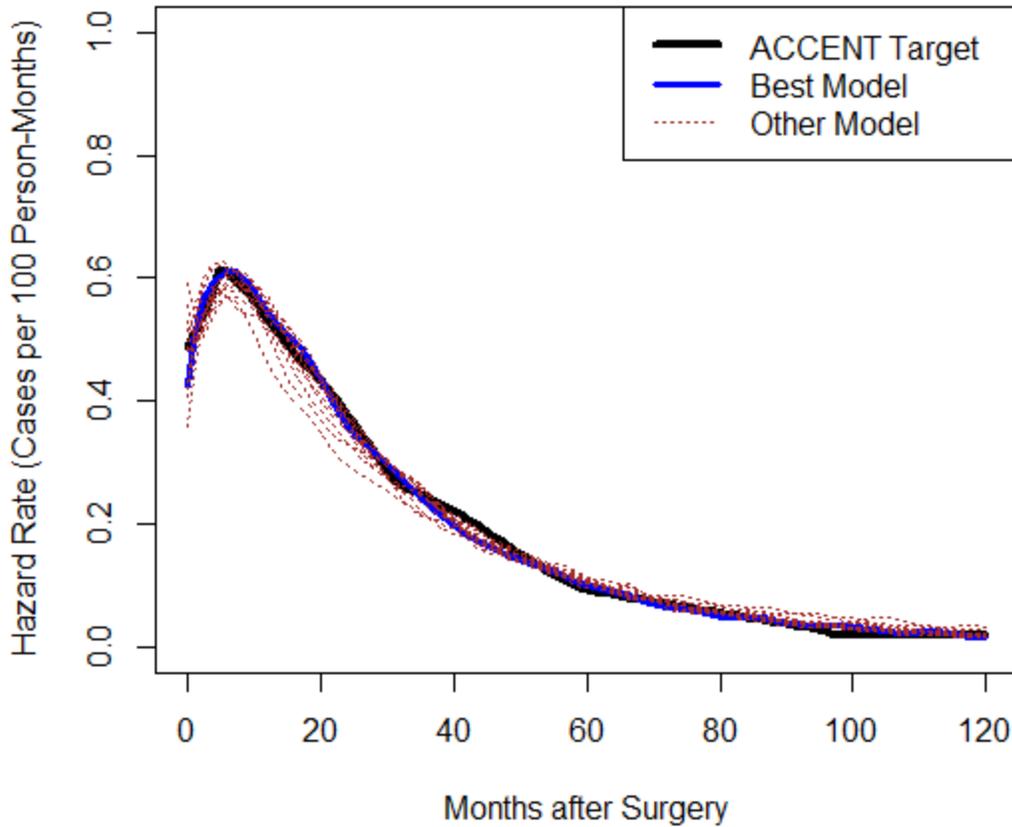


Figure B.9 depicts the calibration results for stage II colon cancer patients treated with surgery alone in the new era. The ‘Best Model’ line is the model-output (smoothed hazard function) using the highest-probability parameter sets (for the time-to-detectable-disease and time-to-clinical-disease processes). The brown dashed lines represent the model-output with the other included parameter sets. The y-axis presents the hazard rate and has been scaled to cases per 100-person months.

The data for this target was assumed to have been generated in the absence of any routine follow-up.

Figure B.10: (Old) Stage II Colon Cancer 5-FU + LV

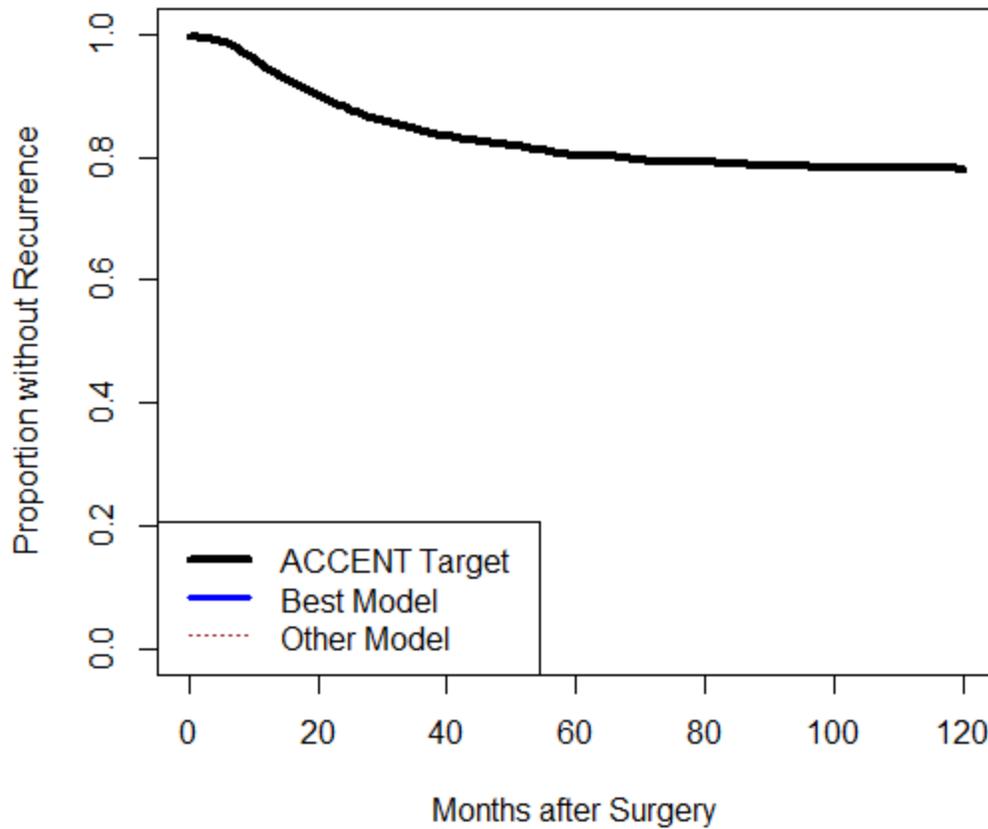


Figure B.10 depicts the calibration results for stage II colon cancer patients treated with 5-FU/LV in the old era. The 'Best Model' line is the model-output (Kaplan-Meier Survival Curve) using the highest-probability parameter sets (for the time-to-detectable-disease and time-to-clinical-disease processes). The brown dashed lines represent the model-output with the other included parameter sets.

The data for this target was assumed to have been generated in the absence of any routine follow-up.

Figure B.11: (New) Stage II Colon Cancer 5-FU + LV

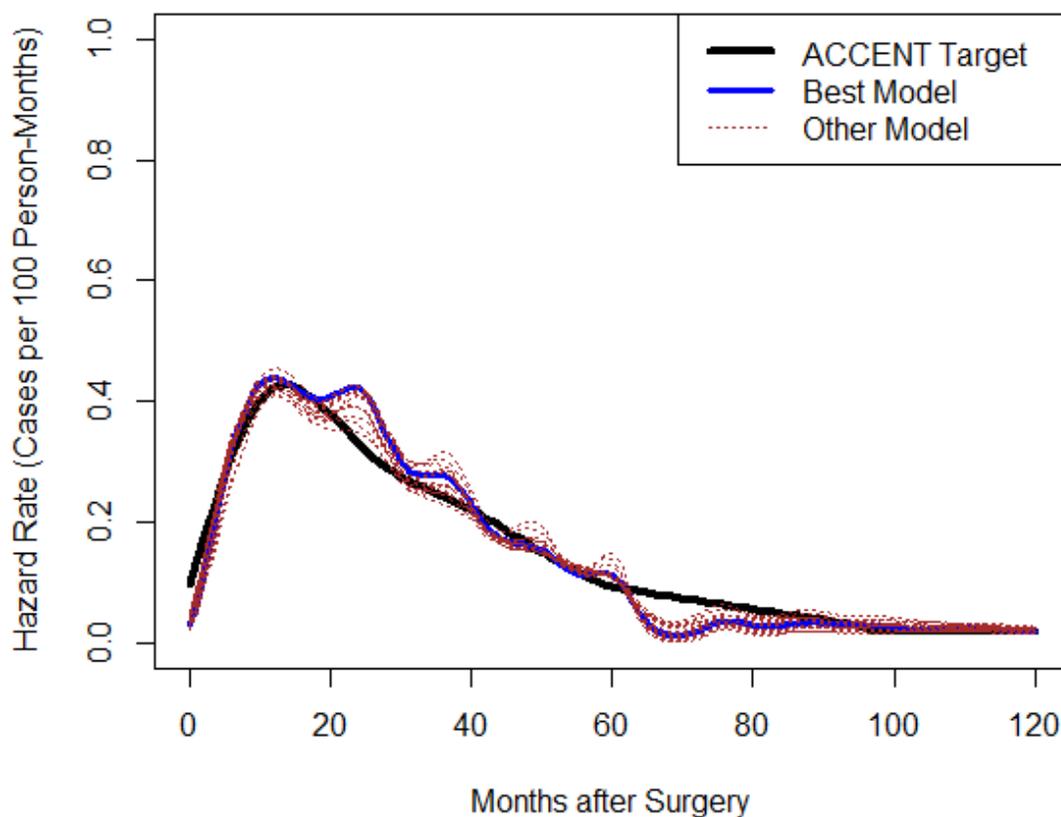


Figure B.11 depicts the calibration results for stage II colon cancer patients treated with adjuvant 5-FU/LV in the new era. The ‘Best Model’ line is the model-output (smoothed hazard function) using the highest-probability parameter sets (for the time-to-detectable-disease and time-to-clinical-disease processes). The brown dashed lines represent the model-output with the other included parameter sets. The y-axis presents the hazard rate and has been scaled to cases per 100-person months.

The data for this target was assumed to have been generated with under the following follow-up schedule: CEA every 3 months for 2 years, then every 6 months for 3 years. CT every year for 5 years.

Figure B.12: Stage II Colon Cancer FOLFOX

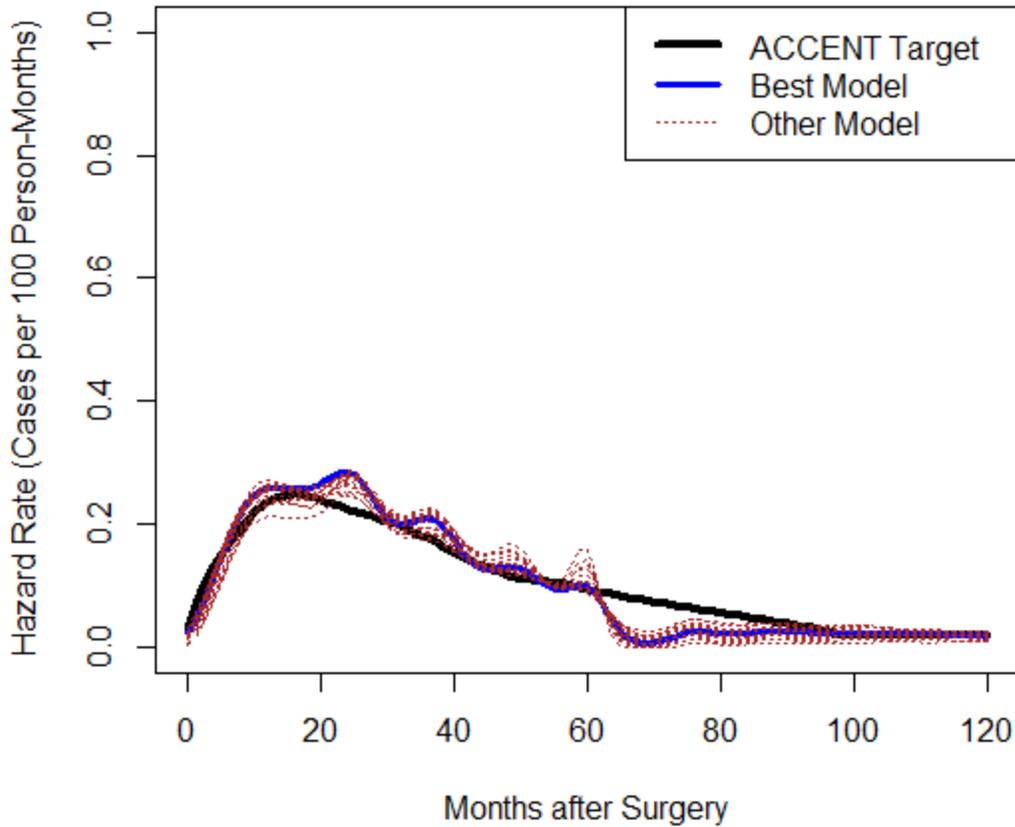


Figure B.12 depicts the calibration results for stage II colon cancer patients treated with adjuvant FOLOFX. The 'Best Model' line is the model-output (smoothed hazard function) using the highest-probability parameter sets (for the time-to-detectable-disease and time-to-clinical-disease processes). The brown dashed lines represent the model-output with the other included parameter sets. The y-axis presents the hazard rate and has been scaled to cases per 100-person months.

The data for this target was assumed to have been generated with under the following follow-up schedule: CEA every 3 months for 2 years, then every 6 months for 3 years. CT every year for 5 years.

Figure B.13: (Old) Stage I Colon Cancer

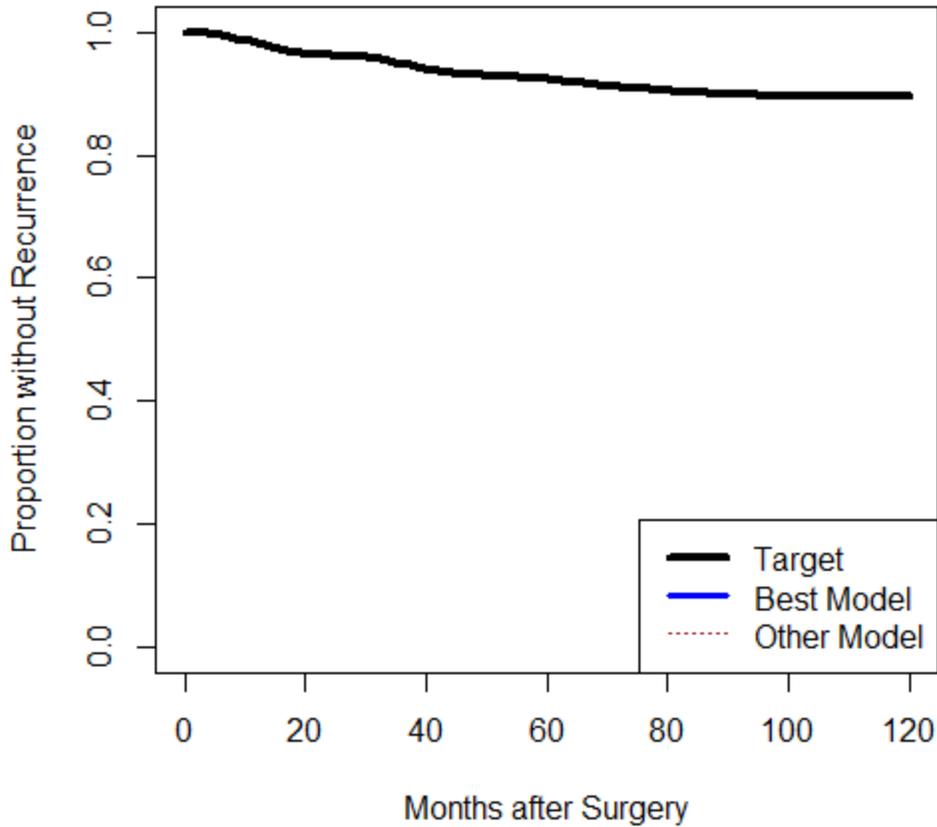


Figure B.13 depicts the calibration results for stage I colon cancer patients treated in the old era. The 'Best Model' line is the model-output (Kaplan-Meier Survival Curve) using the highest-probability parameter sets (for the time-to-detectable-disease and time-to-clinical-disease processes). The brown dashed lines represent the model-output with the other included parameter sets.

The data for this target was assumed to have been generated with under the following follow-up schedule: CEA every 3 months for 2 years, then every 6 months for 3 years. CT scans at 1 and 2 years.

Figure B.14: (New) Stage I Colon Cancer

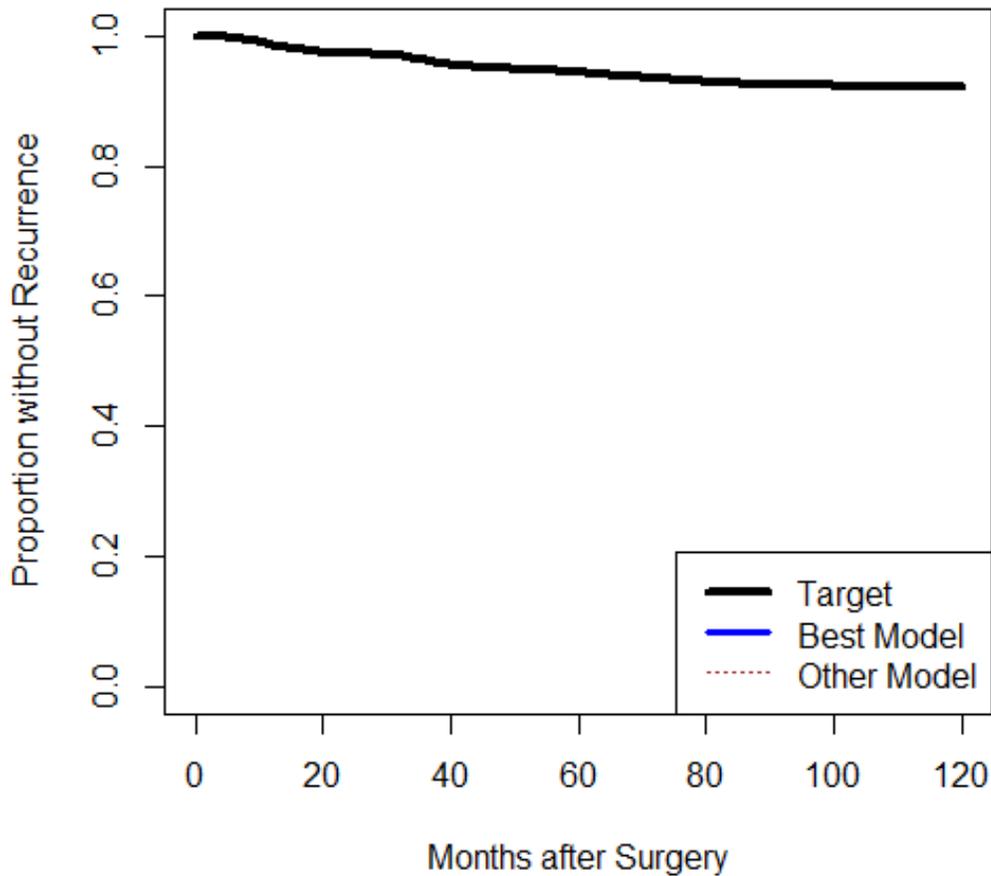


Figure B.14 depicts the calibration results for stage I colon cancer patients treated in the new era. The 'Best Model' line is the model-output (Kaplan-Meier Survival Curve) using the highest-probability parameter sets (for the time-to-detectable-disease and time-to-clinical-disease processes). The brown dashed lines represent the model-output with the other included parameter sets.

The data for this target was assumed to have been generated with under the following follow-up schedule: CEA every 3 months for 2 years, then every 6 months for 3 years. CT scans at 1 and 2 years.

Figure B.15: Stage I Rectal Cancer

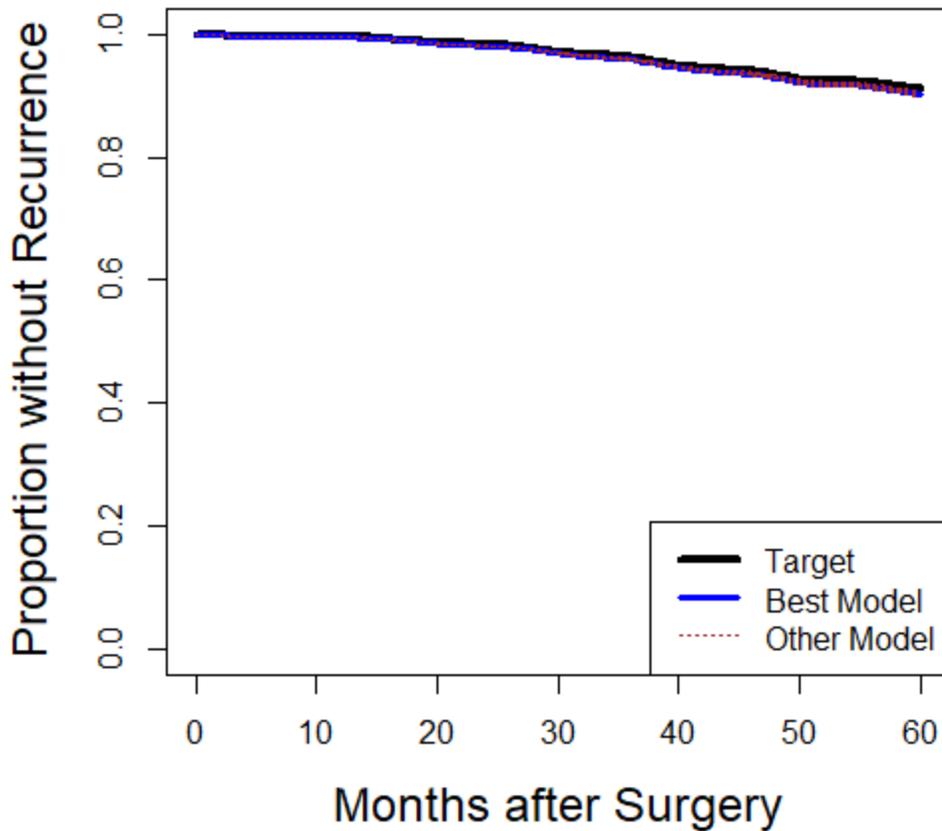


Figure B.15 depicts the calibration results for stage I rectal cancer patients treated with total mesorectal excision surgery. The ‘Best Model’ line is the model-output (Kaplan-Meier Survival Curve) using the highest-probability parameter sets (for the time-to-detectable-disease and time-to-clinical-disease processes). The brown dashed lines represent the model-output with the other included parameter sets.

The data for this target was assumed to have been generated with under the following follow-up schedule: CEA every 3 months for 2 years, then every 6 months for 3 years. CT scans at 1 and 2 years.

Figure B.16: Stage II Rectal Cancer

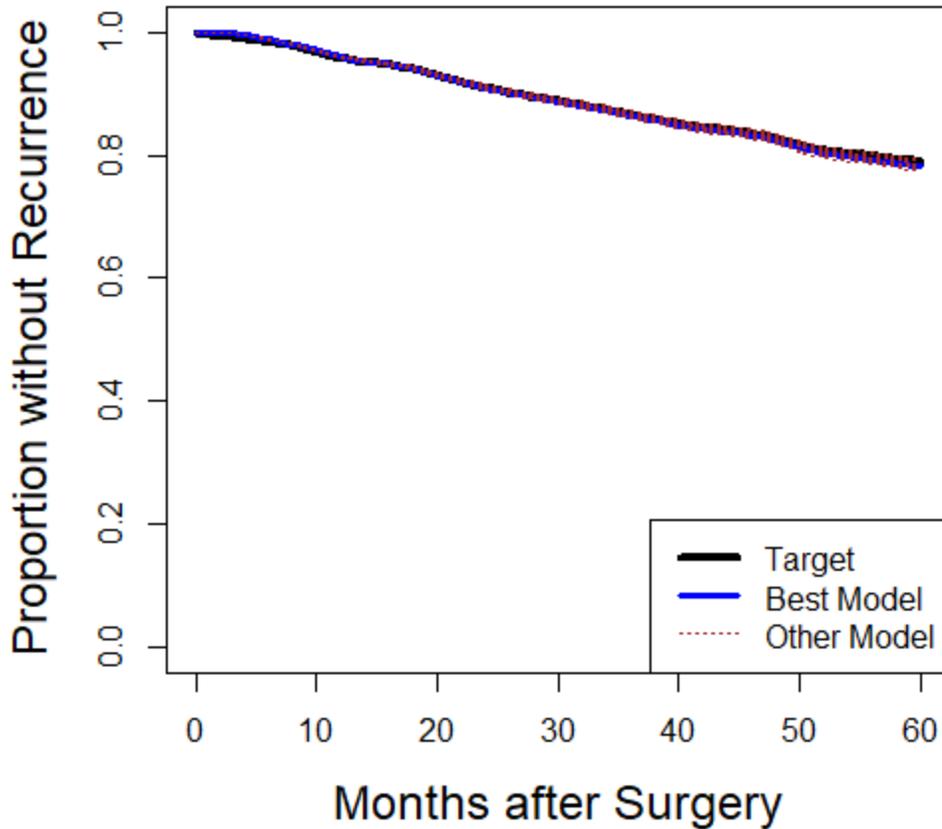


Figure B.16 depicts the calibration results for stage II rectal cancer patients. These patients were treated with total mesorectal excision (TME) surgery and adjuvant 5-FU/LV chemotherapy. Some of them received preoperative or postoperative radiation therapy. The 'Best Model' line is the model-output (Kaplan-Meier Survival Curve) using the highest-probability parameter sets (for the time-to-detectable-disease and time-to-clinical-disease processes). The brown dashed lines represent the model-output with the other included parameter sets.

The target source was a RCT enrolling patients in the late 1990's which did not call for routine surveillance. However, I assumed patients underwent CEA testing every 6 months for 3 years and once a year for 2 years.

Figure B.17: Stage III Rectal Cancer

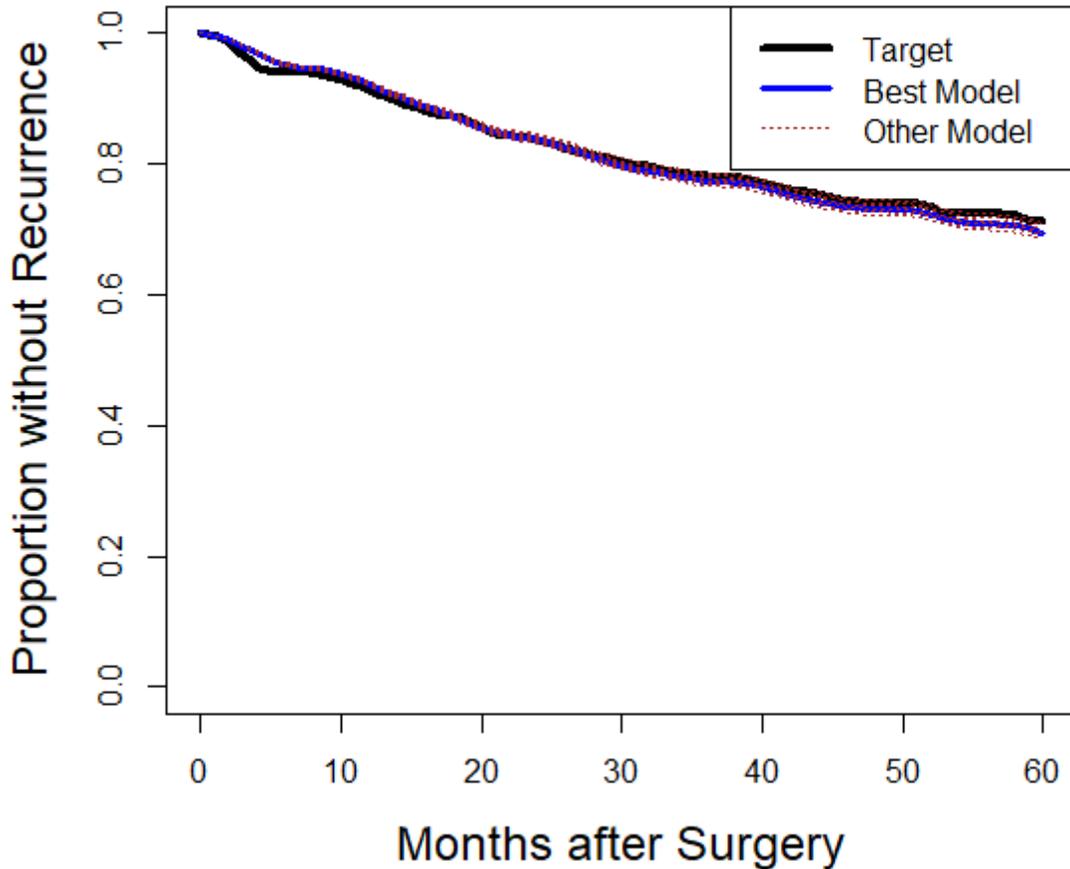


Figure B.17 depicts the calibration results for stage III rectal cancer patients. These patients were treated with total mesorectal excision (TME) surgery, neoadjuvant chemoradiation therapy and adjuvant chemotherapy, both using a FOLFOX regimen. The ‘Best Model’ line is the model-output (Kaplan-Meier Survival Curve) using the highest-probability parameter sets (for the time-to-detectable-disease and time-to-clinical-disease processes). The brown dashed lines represent the model-output with the other included parameter sets.

The data for this target was assumed to have been generated with under the following follow-up schedule: CEA every 6 months for 2 years, then every 12 months for 3 years and annual CT scans.

Table B.1: Surveillance Schedule for the Intensive and Minimal Arms Used in Calibration

Follow-up Arm	Test	3	6	9	12	15	18	21	24	30	36	42	48	54	60
Inten. Arm	CT	.08	.75	.04	.88	0	.42	0	.88	0	.75	0	.75	0	.63
Min. Arm	CT	0	.08	0	.25	0	.25	0	0	0	0	0	0	0	.0
Inten. Arm	CEA	.96	.96	.96	.96	.96	.96	.96	.96	.96	.96	.96	.96	.96	.96
Min. Arm	CEA	.5	.63	.5	.63	.5	.5	.5	.63	.5	.63	.5	.63	.38	.5

Table B.1 depicts the follow-up schedule for the two regimens used to calibrate the natural history model parameters. In particular, these follow-up regimens were used when evaluating model-produced surveillance output in relation to the four meta-analytic targets for the purpose of assigning a probability mass to each pair of hazard functions (time to detectable and clinical disease). The top row gives the time in months (from 3 to 60 months) after the primary resection. In each cell, the number (between 0 and 1) represents the proportion of patients in the simulation that were assigned each test at each time period. These numbers were derived from combining the follow-up regimens for the 8 trials that contributed to the targets depicted in Figure B.3/B.4.

Inten. = Intensive; Min. = Minimal

Table B.2: Description of Distribution of Sojourn Times by Disease Type

DISEASE	Overall Mean Sojourn Time (Range of Means)	Overall Median Sojourn Time (Range of Medians)	Overall 1st Quartile of Sojourn Times (Range of 1st Quartiles)	Overall 3rd Quartile of Sojourn Times (Range of 3rd Quartiles)
Stage III Colon No Adj	17.8 (11-31.9)	10.8 (6.1-20.5)	3.1 (1.5-6.9)	24.6 (15.3-45.1)
Stage III Colon 5-FU + LV	18.3 (10.9-36.7)	12.1 (7.2-25)	4.9 (2.9-10.2)	24.9 (14.7-51.5)
Stage III Colon FOLFOX	21.4 (10.2-39.3)	11.6 (7.3-24.1)	5.3 (3.8-8.8)	28.5 (13.3-63.8)
Stage II Colon No Adj (New)	23.2 (14-36.4)	12.4 (7.2-20.5)	4.0 (2.1-6.8)	30.4 (17.9-51.3)
Stage II Colon No Adj (Old)	24 (14.9-32.6)	12.9 (7.7-18)	4.1 (2.3-5.9)	31.7 (19.2-44.7)
Stage II Colon 5-FU + LV (New)	29.8 (18.2-54.5)	19.8 (11.5-40.6)	7.6 (4.2-16.0)	41.9 (24.6-84.6)
Stage II Colon 5-FU + LV (Old)	23 (15-32.9)	12.2 (7.8-18.3)	3.9 (2.4-6.0)	30.3 (19.3-45)
Stage II Colon FOLFOX	41.5 (23.4-70.0)	29.3 (14.8-59.5)	11.3 (5.6-23.4)	61.8 (31.9-123.1)
Stage I Colon (New)	37.1 (25.1-62.3)	27.8 (17.6-51.5)	11.4 (6.6-22.3)	53.6 (36-97.2)
Stage I Colon (Old)	35.6 (25-48)	26.3 (17.6-36.9)	10.7 (6.7-15.7)	50.9 (35.4-70)
Stage III Rectal	23.4 (11.3-45.3)	12.5 (7.9-27.1)	6.3 (4.8-9.3)	34.5 (16.1-65.3)
Stage II Rectal	30.5 (19.2-56.4)	20.8 (12.3-44.4)	9.6 (6.2-18.0)	49.4 (26.3-86.6)
Stage I Rectal	36.4 (25.2-55.4)	29.2 (17.5-46.9)	13.7 (6.7-22.7)	54.1 (35.9-79.5)

For each combination of disease location, stage, adjuvant therapy, and period, Table B.2 gives information about the distribution of patient sojourn times in months. In particular, the mean, median, and 1st and 3rd quartiles are given. These point estimates represent the model's best prediction of the moment or quantile of the distribution made by averaging over parameter sets. The values in parentheses given below represent the range (across

different parameter sets) of each moment or quantile of the distribution. Though it should be noted that the values are not necessarily equally likely.

Appendix C

Figure C.1: Size of Largest Lesion vs. Proportion of Elapsed Sojourn Time

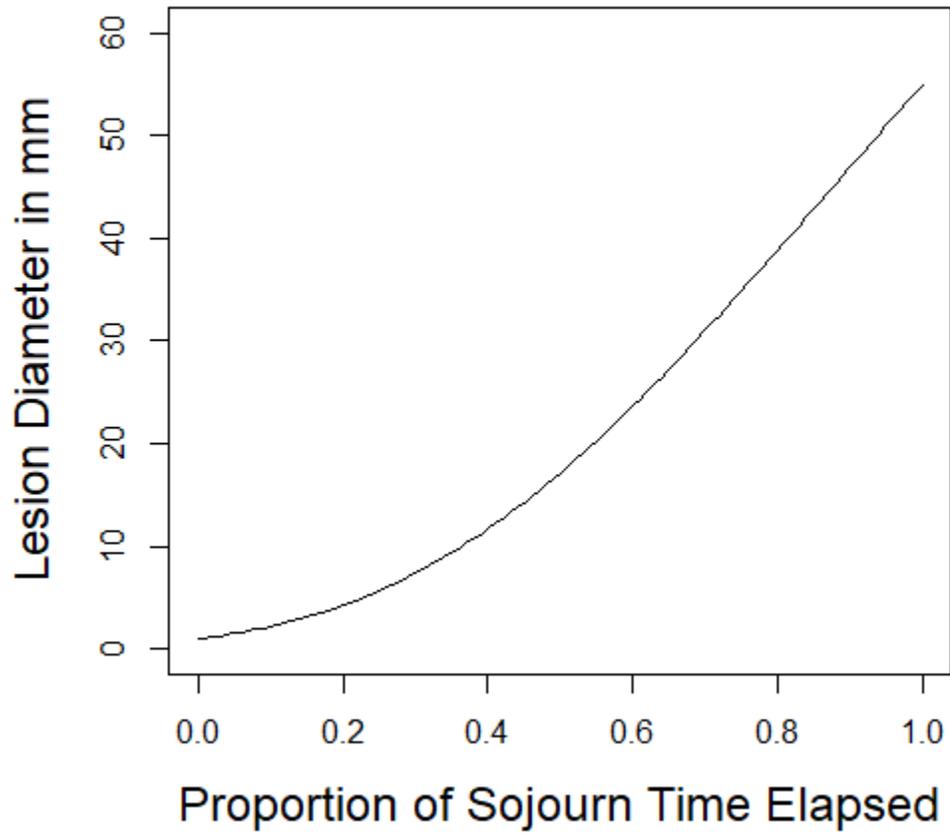
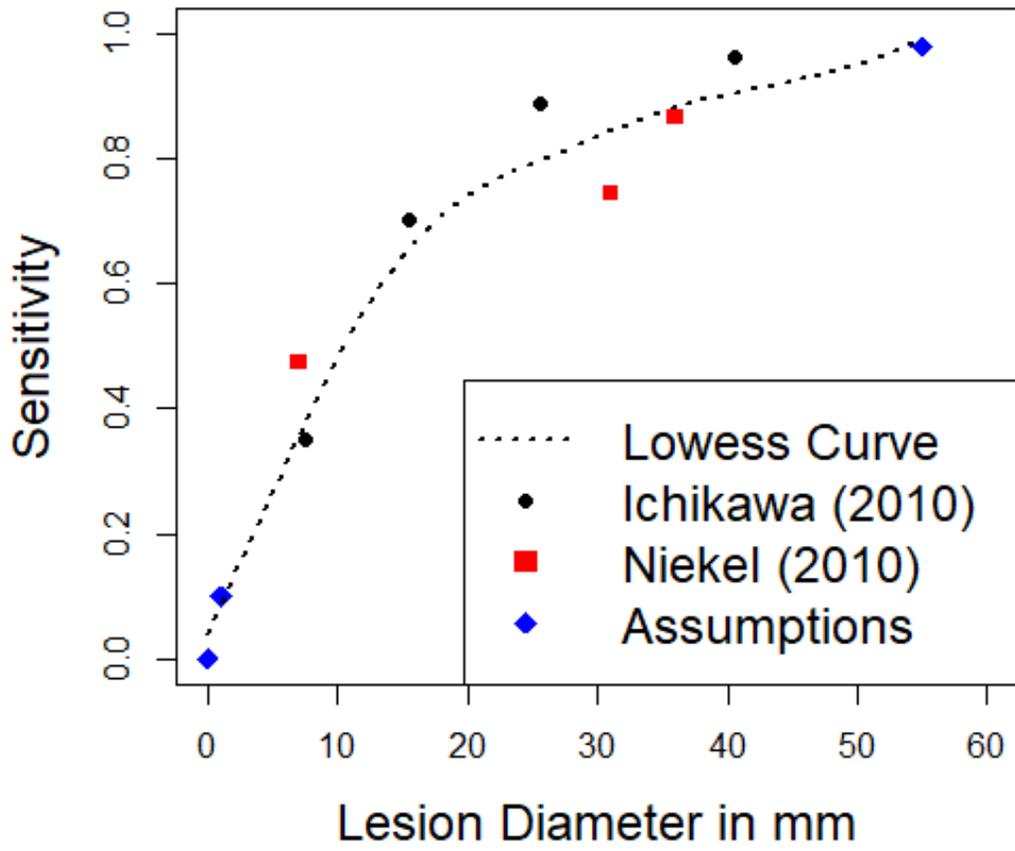


Figure C.2: Sensitivity of CT Scan vs. Diameter of Largest Lesion



Appendix D

Table D1: Costs of Individual Services

Resource	Type of Cost	CPT Codes	Physician Reimbursement or Total Outpatient Reimbursement (2017 US \$)	DRG Codes & DRG Weights	Total Hospital Cost (2017 US \$)	Total Reimbursement (2017 US \$)
FOLLOW-UP & DIAGNOSTIC TESTS						
CT of Pelvis w/ IV	Outpatient	72193	\$229	---	---	\$229
CT of Abdomen w/ IV	Outpatient	74160	\$234	---	---	\$234
CT of Abdomen & Pelvis w/ IV	Outpatient	74177	\$315	---	---	\$315
CT of Chest w/ IV	Outpatient	71260	\$233	---	---	\$233
MRI of Pelvis w/ IV	Outpatient	72196	\$416	---	---	\$416
MRI of Abdomen w/ IV	Outpatient	74182	\$460	---	---	\$460
CEA Assay	Outpatient	82378	\$26	---	---	\$26
Clinical Visit (30 Min)	Outpatient	99203	\$109	---	---	\$109
Diagnostic Laparoscopy	Outpatient	49320	\$338	---	---	\$338
Diagnostic Thoracoscopy w/ biopsy of Nodule	Outpatient	32608	\$394	---	---	\$394
CT-guided FN Biopsy	Outpatient	77012 47000	\$438	---	---	\$438
HEPATIC RESECTION						
No Resection/R2 Resection (Laparotomy)	Inpatient	49000	\$800	407 = 2.0118	\$11,996	\$12,796
Intraoperative Liver US	Inpatient	76998	\$65	Done in 100% Surgery (\$65 * 1)		\$65
Wedge Resection	Inpatient	47100	\$876	405 (MCC) = 5.4464 406 (CC) = 2.7825 407 (None) = 2.0118	405 = \$32,477 406 = \$16,592 407 = \$11,996	
Partial lobectomy	Inpatient	47120	\$2,424			
Trisegmentectomy	Inpatient	47122	\$3,568			
Total left lobectomy	Inpatient	47125	\$3,198			
Total right lobectomy	Inpatient	47130	\$3,436			
Liver Resection (Use)	Inpatient	(0.175, 0.525, 0.1, 0.1, 0.1) \$2,446		(0.44, 0.44, 0.12) \$23,030		\$25,476
Resection-Related Morbidity ¹⁸⁸	Inpatient	---	---	---	---	\$37,425
Resection-Related Mortality ¹⁸⁸	Inpatient	---	---	---	---	\$92,810
PULMONARY RESECTION						
No Resection/R2 Resection (Thoracotomy)	Inpatient	32100	\$842	165 = 1.7898	\$10,672	\$11,514
Mediastinal LN Dissection	Inpatient	38746	\$224	Done in 45% of resections (\$224 * 0.45)		\$101

Wedge Resection 1st Nodule with Thoracoscopy + 2 nd Nodule	Inpatient	32666 32667	\$904 + \$164 = \$1,068	163 (MCC) = 5.0194 164 (CC) = 2.5187 165 (None) = 1.7898	163 = \$29,930 164 = \$15,019 165 = \$10,672	
Segmentectomy	Inpatient	32484	\$1,492			
Lobectomy	Inpatient	32480	\$1,536			
Pneumonectomy	Inpatient	32440	\$1,625			
Bilobectomy	Inpatient	32482	\$1,643			
Lung Resection (Use)	Inpatient	(0.66, 0.07, 0.25, 0.01, 0.01) \$1,226		(0.44, 0.44, 0.12) \$21,058		\$22,284
Resection-Related Morbidity ¹⁸⁹	Inpatient	---	---	---	---	\$43,898
Resection-Related Mortality ¹⁸⁹	Inpatient	---	---	---	---	\$87,796
RESECTION FOR PELVIC RECURRENCE						
No Resection/R2 Resection (Laparotomy)	Inpatient	49000	\$800	735 (F) = 1.2428 708 (M) = 1.3476	\$ 7,754	\$8,554
Excise Sacral Spine Tumor	Inpatient	49215	\$2,316	Done in 32% of resections (\$2,316*0.32)		\$741
IOERT ¹⁹⁰	Inpatient	77425 77469	\$483 \$327	Done in 2/3 of resections (2/3 * \$810)		\$540
Transabdominal Resection: Partial Removal of Rectum	Inpatient	45111	\$1,125	FEMALES 734 (CC) = 2.7192 735 (None) = 1.2428	734 = \$16,214 735 = \$7,411	
Abdomino-Perineal Resection: Removal of Rectum	Inpatient	45110	\$1,926	MALES 707 (CC) = 1.8091 708 (None) = 1.3476	707 = \$10,788 708 = \$8,036	
Pelvic Exenteration	Inpatient	45126	\$2,862			
Pelvic Resection (Use)	Inpatient	(0.23, 0.40, 0.37) \$2,088		(0.88, 0.12) * 0.35 + (0.88, 0.12) * 0.65 \$12,103		\$14,191
Resection-Related Morbidity ¹⁹¹	Inpatient	---	---	---	---	\$30,533
Resection-Related Mortality ¹⁹¹	Inpatient	---	---	---	---	\$61,066
RESECTION OF COLON LOCAL-REGIONAL RECURRENCE IN ABDOMEN						
No Resection/R2 Resection (Laparotomy)	Inpatient	49000	\$800	331 = 1.6491	\$9,833	\$10,633
IOERT ¹⁹⁰	Inpatient	77425 77469	\$483 \$327	Done in 15% of resections (0.15 * \$810)		\$122
Abdominal Resection for LR Recurrence in Colon Cancer ¹⁶⁶	Inpatient	---	---	---	---	\$45,239
(NEO)ADJUVANT THERAPY FOR RESECTABLE DISEASE						
3 Months of FOLFOX ¹⁹²	Drug	---	---	---	---	\$16,414
3 Months of 5- FU/LV ¹⁹²	Drug	---	---	---	---	\$3,679
(Neo)adjuvant External-Beam	Outpatient	---	---	---	---	\$6,400

Radiation Therapy ¹⁶⁸						
-------------------------------------	--	--	--	--	--	--

Table D.1 gives the costs or 2017 Medicare reimbursement rates for individual services. The first section gives the values for follow-up and diagnostic tests. These procedures are generally outpatient procedures. The third column identifies relevant CPT codes, and the 4th column gives the total costs (professional + technical component). The next four sections break down individual services for salvage treatment for hepatic, pulmonary, pelvic, and abdominal recurrence, respectively. The last of these relies mostly on an estimate from the literature. The other three sections detail CPT codes and reimbursement levels for physician fees (column 4). They do not include a technical component. In each of the 3 types of salvage surgeries for which a semi-micro-costing approach was taken, there were several CPT codes that might be used. For example, for hepatic metastectomy, there were 5 different common types of surgery found in the literature. While these 5 different procedures have different physician reimbursement rates, they share a common DRG and thus reimbursement for inpatient services. Thus variation in the procedure is assumed only to affect the physician fees. For Hepatic and Pulmonary metastectomy, there were 3 levels of each type of DRG: one for major comorbidities or complications (MCC), one for comorbidities or complications (CC), and one for none of these (None). Columns 5 and 6 give the DRGs and payment weights and reimbursement rates for each DRG, respectively, using a base payment rate of \$5,962.93 per unit of weight. In the case of pelvic salvage, the DRG varies depending upon the sex of the patient. Moreover, only two levels were available for each DRG (CC vs None). In each case, the different procedures and levels of DRG are combined via a weighted average based on weights taken from the literature. These values are given in the row immediately under the set of rows identifying the different procedures and DRGs. These summary rows are identified in the first column by the type of salvage followed by '(Use)'. For example, 'Liver Resection (Use)'. The last column of Table D.1 gives the total costs, combining both physician fees and inpatient costs. For each of the three-mentioned types of salvage surgery, the rows immediately below this summary row give the average cost per case of treatment-related morbidity and mortality, respectively. In the case of surgery for local-regional recurrence in the abdomen (for colon cancer patients only), the mean cost per case of treatment is taken from the literature. This includes the expected costs associated with morbidity and mortality. For all 4 types of salvage surgery, the first row identifies the CPT codes, physician reimbursement rates, the selected level of DRG – None in every case, inpatient reimbursement rates, and total costs associated with failed resection or incomplete (R2) resection. Finally, the last section gives the costs for neoadjuvant and adjuvant radiation therapy and chemotherapy. In the latter case, costs are given per 3 months of treatment. These values were taken from the literature. The assumed utilization rates for these are given in Table D.2, and they are aggregated with salvage surgery costs in Table D.3.

Table D.2: Important Utilization Assumptions Regarding Salvage Surgery

TYPE of RESOURCE	SALVAGE SITE	ASSUMPTIONS
Distribution of DRGs	Liver/Lung Surgery ¹⁵³	44% MCC, 44% CC, 12% None
	Pelvic Surgery ^{67 153}	88% CC, 12% None 65% Male, 35% Female
Distribution of CPT Codes	Liver Surgery ¹⁹³	70% partial (<3 segments), 30% major Among partial: 75% Partial Lobectomy, 25% Wedge Among major: 1/3 Total Left, 1/3 Total Right, 1/3 Trisegmentectomy
	Lung Surgery ^{194 195}	66% Wedge (2 lesions), 25% lobectomy, 7% Segmentectomy, 1% Pneumonectomy, 1% Bilobectomy
	Pelvic Surgery ^{8 10 11 67}	23% Transabdominal Resection, 40% Abdominal-Perineal Resection, 37% Pelvic Exenteration
Cumulative Incidence of 30-day Treatment-Related Mortality and Morbidity	Liver Surgery ^{65 188 193}	Mortality = 2.5% Morbidity = 20%
	Lung Surgery ^{66 196 197}	Mortality = 0.9% Morbidity = 17.5%
	Pelvic Surgery ^{8 61}	Mortality = 2.2% Morbidity = 17%
Additional Inpatient Services	Liver Surgery ^{153 198}	100% get Intraoperative Ultrasound
	Lung Surgery ^{90 194}	45% get Mediastinal Lymph Node Dissection
	Pelvic Surgery ^{8 67}	31% get Removal of Sacral Tumor
Utilization of Chemotherapy	Liver Surgery ^{193 199}	16% neoadjuvant, 28% adjuvant, 19% both; 90% double-agent, 10% single agent; Neoadjuvant = 3 months, Adjuvant = 6 months, Both = 3 + 3 months
	Lung Surgery ^{90 194 195}	56% adjuvant; 90% double-agent, 10% single-agent; Adjuvant = 6 months
	Liver + Lung Surgery ¹⁶⁹	55% neoadjuvant; 60% double-agent, 40% single-agent Neoadjuvant = 3 months 50% adjuvant; 50% double-agent, 50% single-agent Adjuvant = 3 months
	Pelvic Surgery ⁸	33% neoadjuvant; 100% single-agent Neoadjuvant = 1 month
	Local-Regional Abdominal Surgery ²⁰⁰	20% neoadjuvant; 100% single-agent; Neoadjuvant = 1 month; 60% adjuvant; 90% double-agent, 10% single-agent; Adjuvant = 6 months
Utilization of Radiation Therapy	Pelvic Surgery ^{8 10}	33% neoadjuvant EBRT 67% IOERT

	Local-Regional Abdominal Surgery ²⁰⁰	25% neoadjuvant or adjuvant EBRT 15% IOERT
Re-Salvage Rates	Liver Surgery ¹⁷¹	14%
	Lung Surgery ^{90 197 201}	20%
	Liver + Lung Surgery	10%
	Pelvic Surgery ¹¹	10%
	Local-Regional Abdominal Surgery ²⁰⁰	25%
Distribution of Salvage Sites	Colon ⁴	Liver = 60%, Lung = 20%, Liver + Lung = 5%, LR = 15%
	Rectal ⁴	Liver = 40%, Lung = 30%, Liver + Lung = 5%, LR = 25%

Table D.2 identifies important assumptions (and their sources in the literature) regarding utilization of resources for the different types of salvage treatment. These assumption were necessary to translate the costs associated with individual services given in Table D.1 into the aggregate cost estimates given in Table D.3. The terms ‘double-agent’ and ‘single-agent’ refer to the number of chemotherapy agents used (2 or 1, respectively). In the model, I use FOLFOX for all double-agent regimens and 5-FU/LV for all single-agent regimens. EBRT and IOERT stand for external-beam radiation therapy (outpatient radiation therapy) and intraoperative electron-beam radiation therapy (given during surgery). MCC, CC, and None stand for major complication or comorbidity, complication or comorbidity, and no complication or comorbidity, respectively.

Table D.3: Aggregate Costs of Surveillance and Salvage and Palliative Treatment

AGGREGATE RESOURCE CATEGORY	COST (2017 US \$)	CALCULATIONS/NOTES
FOLLOW-UP TESTS		
Full CT Scan for Rectal Patients	\$548	\$315 + \$233
Full CT Scan for Colon Patients	\$467	\$234 + \$233
Cost of a True Positive CT: Colon	\$40	$\left(\frac{338 + 394}{2}\right) \times 0.05 + (\$438 \times 0.05)$
Cost of a True Positive CT: Rectal	\$144	$(\$416 \times 0.25) + \left(\frac{338 + 394}{2}\right) \times 0.05 + (\$438 \times 0.05)$
Cost of a False Positive CT: Colon	\$374	$\left(\frac{234 + 233}{2}\right) + (460) + (438) + \left(\frac{338 + 394}{2}\right)$
Cost of a False Positive CT: Rectal	\$369	$\left(\frac{234 + 233}{2}\right) + \left(\frac{416 + 460}{2}\right) + (438) + \left(\frac{338 + 394}{2}\right)$
TREATMENT WITH CURATIVE INTENT		
Cost of Failed/R2 Resection: (Colon)	\$12,215	$(\$12,796 \times 0.65) + (\$11,514 \times 0.2) + (\$10,633 \times 0.15)$
Cost of Failed/R2 Resection: (Rectal)	\$11,287	$(\$12,796 \times 0.4) + (\$11,514 \times 0.35) + (\$8,554 \times 0.25)$
Expected Cost of Hepatic Resection: Surgery	\$35,346	$(\$25,476 + \$65) + (\$37,425 \times 0.2) + (\$92,810 \times 0.025)$
Expected Cost of Hepatic Resection: (Neo)Adjuvant Therapy	\$16,655	$0.16 \times ((0.9 \times \$16,414) + (0.1 \times \$3,679)) + 0.47 \times 2 \times ((0.9 \times \$16,414) + (0.1 \times \$3,679))$
Total Expected Cost of Hepatic Salvage Treatment	\$52,001	\$35,346 + \$16,655
Expected Cost of Pulmonary Resection: Surgery	\$30,857	$(\$22,284 + \$101) + (\$43,898 \times 0.175) + (\$87,796 \times 0.009)$
Expected Cost of Pulmonary Resection: (Neo)Adjuvant Therapy	\$16,957	$0.56 \times 2 \times ((0.9 \times \$16,414) + (0.1 \times \$3,679))$
Total Expected Cost of Pulmonary Salvage Treatment	\$47,814	\$30,857 + \$16,957
Expected Cost of Staged Hepatic & Pulmonary Resection: Surgery	\$66,203	\$35,346 + \$30,857
Expected Cost of Staged Hepatic & Pulmonary Resection:	\$11,249	$0.55 \times (0.6 \times 16414 + 0.4 \times 3679) + 0.50 \times (0.5 \times 16414 + 0.5 \times 3679)$

(Neo)Adjuvant Therapy		
Total Expected Cost of Staged Hepatic & Pulmonary Salvage Treatment	\$77,452	\$66,203 + \$11,249
Expected Cost of Pelvic Resection: Surgery	\$22,159	$(\$14,191 + \$74 + \$540) + (\$30,533 \times 0.175) + (\$61,066 \times 0.022)$
Expected Cost of Pelvic Resection: (Neo)Adjuvant Therapy	\$ 2,542	$1/3 \times (\$6,400 + \frac{\$3,679}{3})$
Total Expected Cost of Pelvic Salvage Treatment	\$24,701	\$22,159 + \$2,542
Expected Cost of Abdominal Resection of Colon LR Recurrence: Surgery	\$45,361	\$45,239 + \$122
Expected Cost of Abdominal Resection of Colon LR Recurrence: (Neo)Adjuvant Therapy	\$20,014	$0.2 \times (\frac{\$3,679}{3}) + 0.25 * \$6,400 + 0.6 * 2 * (0.9 * \$16,414 + 0.1 * \$3,679)$
Total Expected Cost of Abdominal Salvage Treatment of Colon LR Recurrence	\$65,375	\$45,361 + \$20,014
Expected Cost of R0-1 Surgical Resection + (Neo)Adjuvant Therapy (Colon)	\$54,442	$(\$52,001 \times 0.6) + (\$47,814 \times 0.2) + (\$77,452 \times 0.05) + (\$65,375 \times 0.15)$
Expected Cost of R0-1 Surgical Resection + (Neo)Adjuvant Therapy (Rectal)	\$45,192	$(\$52,001 \times 0.4) + (\$47,814 \times 0.3) + (\$77,452 \times 0.05) + (\$24,701 \times 0.25)$
ADDITIONAL DOWNSTREAM COSTS FOR PATIENTS TREATED WITH CURATIVE INTENT		
Expected Cost per Round of CT Follow-up (Colon)	\$486	Includes Expected False-Positive Costs. Unless patient is in terminal phase (last 9 months), applies every 6 months for 2 years and then annually for 3 years.
Expected Cost per Round of CT Follow-up (Rectal)	\$567	Includes Expected False-Positive Costs. Unless patient is in terminal phase (last 9 months), applies every 6 months for 2 years and then annually for 3 years
Expected Cost of Possible Second Surgical Resection with Curative Intent + (Neo)Adjuvant Therapy (Colon)	\$9,255	$\$54,442 \times 0.17$
Expected Cost of Possible Second Surgical Resection	\$6,779	$\$45,192 \times 0.15$

with Curative Intent + (Neo)Adjuvant Therapy (Rectal)		
Age-Specific Monthly Costs for Patient Care ¹⁶⁴	< 75 = \$1,093 75 – 84 = \$656 ≥ 85 = \$473	Applies starting in 7 th month after the date of salvage surgery until the last 9 months of the patient’s life (in the case of cancer-death) or in all remaining lifetime until 10 years (if no cancer death).
Monthly Costs for Terminal Cancer ¹⁶⁴	\$4,091	Applies for last 9 months of patient’s life in the cast of cancer death. If patient lives 9 or fewer months, only this cost is applied, not patient-care.
PALLIATIVELY-TREATED PATIENTS		
Monthly Costs for Terminal Cancer ¹⁶⁴	\$4,091	Applies for last 9 months of patient’s life in the cast of cancer death. If patient lives 9 or fewer months, only this cost is applied, not patient-care or initial treatment costs.
Monthly Costs for Initial Treatment ¹⁶⁴	\$7,438	Applies from the time of diagnosis of unresectable recurrent disease up to at most the next 8 months. However, applies only if patient lives at least 10 months because terminal disease gets priority.
Age-Specific Monthly Costs for Patient Care ¹⁶⁴	< 75 = \$3,683 75 – 84 = \$2,210 ≥ 85 = \$1,473	Applies to any month patient is alive that is not attributed a terminal-cancer or initial-treatment cost.

Table D.3 shows the aggregate costs for surveillance, salvage treatment, and palliative treatment. The last column shows the calculations used to derive the costs or provides notes about the use of the value. The first section shows the aggregate surveillance costs. The second section builds up to the expected cost of salvage treatment for rectal and colon cancer patients. The third section gives additional aggregate costs associated with salvage treatment over and above the initial treatment. The last section gives costs associated with palliative treatment. LR Recurrence = local-regional recurrence.

Table D.4: Incremental Increases in Life-Expectancy (Days) by Cohort

Cohort	Discount Rate	No Surveillance: Life-Expectancy (Years)	1 CT Scan (Days)	CEA + 1 CT Scan (Days)	CEA + 2 CT Scans (Days)	CEA + 3 CT Scans (Days)	CEA + 5 CT Scans (Days)
Stage III Colon 60	0%	16.342 (16.334-16.350)	48.9 (48.2-49.6)	87.6 (86.5-88.7)	9.5 (8.8-9.9)	4.4 (4.0-4.7)	3.6 (3.3-4.4)
Stage III Colon 60	3%	11.733 (11.728-11.738)	30.3 (29.9-30.7)	53.3 (52.6-53.7)	5.5 (5.1-5.8)	2.2 (1.8-2.6)	1.8 (1.5-2.2)
Stage III Rectal 60	0%	17.167 (17.159-17.175)	40.9 (40.1-41.6)	74.8 (73.7-75.9)	8.0 (7.3-8.4)	2.9 (2.2-3.3)	4.0 (3.6-4.4)
Stage III Rectal 60	3%	12.263 (12.258-12.268)	25.6 (24.8-25.9)	46.0 (45.3-46.4)	4.7 (4.4-5.1)	1.1 (0.7-1.5)	1.8 (1.5-2.2)
Stage II Colon 60	0%	19.602 (19.595-19.610)	46.0 (45.6-46.7)	79.9 (78.8-81.0)	8.0 (7.7-8.8)	4.0 (3.6-4.4)	4.0 (3.6-4.4)
Stage II Colon 60	3%	13.818 (13.813-13.822)	29.2 (28.5-29.6)	49.3 (48.9-50.0)	5.1 (4.7-5.1)	2.2 (1.8-2.6)	1.8 (1.8-2.2)
Stage II Rectal 60	0%	18.980 (18.972-18.988)	49.6 (48.9-50.7)	103.7 (102.6-104.8)	10.6 (9.9-10.9)	7.7 (6.9-8.0)	9.5 (8.8-9.9)
Stage II Rectal 60	3%	13.459 (13.454-13.463)	31.4 (30.7-31.8)	63.5 (63.1-64.2)	6.2 (6.2-6.6)	4.4 (4.0-4.7)	5.1 (4.7-5.5)
Stage III Colon 65	0%	13.453 (13.448-13.458)	40.9 (40.1-41.2)	71.5 (70.8-72.3)	7.3 (6.9-7.7)	2.6 (2.2-2.9)	2.6 (2.2-2.9)
Stage III Colon 65	3%	10.159 (10.156-10.163)	26.6 (26.3-27.0)	46.0 (45.6-46.7)	4.4 (4.4-4.7)	1.1 (0.7-1.5)	1.5 (1.1-1.5)
Stage III Rectal 65	0%	14.422 (14.417-14.427)	31.0 (30.7-31.4)	58.4 (58.0-59.1)	5.8 (5.5-6.2)	1.8 (1.5-2.2)	1.8 (1.5-2.2)
Stage III Rectal 65	3%	10.822 (10.819-10.825)	20.4 (20.1-20.8)	38.0 (37.6-38.3)	3.6 (3.3-3.6)	0.7 (0.7-1.1)	0.7 (0.4-0.7)
Stage II Colon 65	0%	16.333 (16.328-16.338)	36.1 (35.8-36.5)	61.0 (60.2-61.3)	5.8 (5.8-6.2)	2.6 (2.2-2.9)	2.2 (1.8-2.6)
Stage II Colon 65	3%	12.109 (12.106-12.112)	24.1 (23.7-24.5)	39.8 (39.4-40.1)	3.6 (3.6-4.0)	1.5 (1.1-1.5)	1.1 (1.1-1.5)
Stage II Rectal 65	0%	15.895 (15.89-15.90)	38.7 (38.3-39.1)	78.1 (77.7-78.8)	7.3 (6.9-7.3)	5.5 (5.1-5.8)	6.2 (5.8-6.2)
Stage II Rectal 65	3%	11.845 (11.842-11.848)	25.6 (25.2-25.9)	50.4 (50.0-51.1)	4.4 (4.4-4.7)	3.3 (2.9-3.6)	3.3 (3.3-3.6)
Stage III Colon 70	0%	10.858 (10.854-10.862)	34.7 (34.3-35.0)	58.4 (57.7-59.1)	5.8 (5.5-6.2)	1.8 (1.5-1.8)	1.5 (1.1-1.5)

Stage III Colon 70	3%	8.598 (8.595-8.600)	24.1 (23.7- 24.5)	40.5 (40.1- 40.9)	4.0 (3.6-4.0)	0.7 (0.7-1.1)	0.7 (0.4-0.7)
Stage III Rectal 70	0%	11.865 (11.861-11.869)	24.5 (24.1- 24.8)	44.2 (43.4- 44.5)	4.4 (4.0-4.7)	1.5 (1.1-1.5)	1.1 (0.7-1.5)
Stage III Rectal 70	3%	9.328 (9.325-9.331)	16.8 (16.8- 17.2)	30.3 (29.9- 30.7)	2.9 (2.6-2.9)	0.7 (0.4-0.7)	0.4 (0.4-0.7)
Stage II Colon 70	0%	13.229 (13.225-13.233)	28.8 (28.5- 29.2)	47.1 (46.7- 47.8)	4.7 (4.4-4.7)	1.8 (1.5-2.2)	1.5 (1.1-1.5)
Stage II Colon 70	3%	10.291 (10.289-10.294)	20.4 (20.1- 20.8)	32.9 (32.5- 33.2)	2.9 (2.9-3.3)	1.1 (1.1-1.5)	0.7 (0.4-0.7)
Stage II Rectal 70	0%	12.997 (12.993-13.001)	29.2 (28.8- 29.6)	58.0 (57.3- 58.4)	5.8 (5.5-6.2)	3.6 (3.3-3.6)	3.6 (3.6-4.0)
Stage II Rectal 70	3%	10.153 (10.150-10.156)	20.4 (20.1- 20.8)	39.8 (39.4- 40.1)	4.0 (3.6-4.0)	2.2 (2.2-2.6)	2.2 (2.2-2.6)
Stage III Colon 75	0%	8.243 (8.239-8.246)	29.9 (29.6- 30.3)	48.5 (47.8- 48.9)	4.0 (4.0-4.4)	0.7 (0.7-1.1)	0.0 (-0.4- 0.4)
Stage III Colon 75	3%	6.844 (6.841-6.846)	21.9 (21.5- 22.3)	36.1 (35.8- 36.5)	2.9 (2.6-2.9)	0.4 (0.4-0.7)	-0.4 (-0.4- 0.0)
Stage III Rectal 75	0%	6.758 (6.755-6.761)	42.0 (41.6- 42.7)	59.9 (59.5- 60.6)	2.6 (2.6-2.9)	0.4 (0.4-0.7)	0.0 (0.0-0.4)
Stage III Rectal 75	3%	5.69 (5.687-5.692)	31.0 (30.7- 31.0)	44.5 (44.2- 44.9)	1.8 (1.8-2.2)	0.0 (0.0-0.4)	0.0 (-0.4- 0.0)
Stage II Colon 75	0%	10.315 (10.312-10.319)	23.0 (22.6- 23.4)	35.8 (35.4- 36.1)	3.3 (2.9-3.3)	1.1 (0.7-1.1)	0.7 (0.4-0.7)
Stage II Colon 75	3%	8.411 (8.408-8.413)	17.2 (16.8- 17.2)	26.6 (26.3- 27.0)	2.2 (2.2-2.6)	0.7 (0.4-0.7)	0.4 (0.4-0.7)
Stage II Rectal 75	0%	10.156 (10.153-10.160)	24.8 (24.5- 25.2)	37.2 (36.9- 37.6)	3.6 (3.3-3.6)	2.2 (1.8-2.2)	1.8 (1.8-2.2)
Stage II Rectal 75	3%	8.286 (8.283-8.288)	18.6 (18.2- 19.0)	27.7 (27.4- 28.1)	2.6 (2.2-2.6)	1.5 (1.1-1.5)	1.1 (1.1-1.1)

Table D.4 contains the results of the efficacy analyses. For each cohort, the third column gives the baseline life-expectancy (LE) in years (discounted at 0% and 3%). By baseline, I mean in the absence of any routine surveillance. The next five columns give the incremental increases in LE in days comparing the surveillance-schedule identified by the column title, e.g., ‘CEA + 2 CT Scans’, with the previous regimen (the one to the left). In each cell, the first entry is the point estimate, and 95% confidence intervals are given in the second row in parentheses. These quantify the precision of the point estimate with respect to Monte Carlo error.

Table D.5: Incremental Increases in Mean Total Costs (2017 US \$) by Cohort

Cohort	Discount Rate	<i>Ex-Ante</i> Expected Costs of Recurrence & Surveillance with No Surveillance	1 CT Scan	CEA + 1 CT Scan	CEA + 2 CT Scans	CEA + 3 CT Scans	CEA + 5 CT Scans
Stage III Colon 60	0%	35,034 (34,988-35,079)	3,003 (2,961- 3,046)	6,222 (6,179- 6,264)	757 (741- 773)	614 (598- 630)	1,009 (993- 1,025)
Stage III Colon 60	3%	30,623 (30,584-30,662)	2,881 (2,846- 2,917)	5,792 (5,757- 5,828)	745 (732- 759)	583 (569- 596)	901 (888-915)
Stage III Rectal 60	0%	30,642 (30,599-30,685)	2,244 (2,205- 2,283)	4,593 (4,555- 4,632)	716 (701- 731)	605 (590- 620)	1,120 (1,105- 1,135)
Stage III Rectal 60	3%	26,745 (26,708-26,782)	2,155 (2,122- 2,188)	4,286 (4,253- 4,318)	706 (694- 719)	578 (566- 591)	998 (986- 1,011)
Stage II Colon 60	0%	17,706 (17,671-17,740)	2,259 (2,228- 2,289)	4,467 (4,436- 4,498)	674 (662- 685)	605 (594- 617)	1,108 (1,096- 1,119)
Stage II Colon 60	3%	15,397 (15,368-15,427)	2,140 (2,115- 2,166)	4,099 (4,073- 4,125)	655 (645- 665)	566 (556- 576)	977 (967-986)
Stage II Rectal 60	0%	22,082 (22,044-22,120)	2,173 (2,139- 2,206)	5,180 (5,147- 5,214)	749 (737- 760)	741 (729- 752)	1,389 (1,377- 1,400)
Stage II Rectal 60	3%	18,727 (18,696-18,759)	2,065 (2,037- 2,092)	4,721 (4,693- 4,749)	727 (717- 736)	693 (683- 702)	1,224 (1,215- 1,234)
Stage III Colon 65	0%	35,580 (35,547-35,612)	3,112 (3,082- 3,142)	6,246 (6,216- 6,275)	783 (769- 797)	602 (588- 616)	1,024 (1,010- 1,038)
Stage III Colon 65	3%	31,404 (31,376-31,432)	2,994 (2,968- 3,020)	5,826 (5,801- 5,852)	771 (759- 783)	574 (562- 586)	912 (900-925)
Stage III Rectal 65	0%	29,394 (29,364-29,424)	2,177 (2,150- 2,204)	4,590 (4,563- 4,617)	719 (707- 730)	622 (611- 634)	1,089 (1,078- 1,101)
Stage III Rectal 65	3%	25,809 (25,783-25,835)	2,100 (2,077- 2,123)	4,285 (4,263- 4,308)	709 (699- 719)	591 (582- 601)	972 (962-981)
Stage II Colon 65	0%	16,946 (16,922-16,971)	2,194 (2,173- 2,215)	4,300 (4,278- 4,321)	668 (659- 677)	588 (579- 597)	1,066 (1,057- 1,075)
Stage II Colon 65	3%	14,840 (14,819-14,861)	2,083 (2,065- 2,100)	3,958 (3,940- 3,977)	649 (641- 656)	550 (542- 557)	941 (934-949)
Stage II Rectal 65	0%	20,775 (20,749-20,801)	2,166 (2,144- 2,188)	4,993 (4,970- 50,16)	755 (745- 765)	756 (746- 766)	1,399 (1,389- 1,409)

Stage II Rectal 65	3%	17,744 (17,722-17,766)	2,056 (2,037- 2,075)	4,576 (4,557- 4,595)	735 (727- 743)	708 (700- 716)	1,239 (1,231- 1,248)
Stage III Colon 70	0%	31,786 (31,758-31,814)	3,462 (3,436- 3,487)	6,388 (6,362- 6,413)	853 (842- 864)	648 (638- 659)	909 (899-920)
Stage III Colon 70	3%	28,374 (28,349-28,399)	3,321 (3,299- 3,344)	5,979 (5,957- 6,001)	834 (825- 843)	611 (602- 620)	810 (801-819)
Stage III Rectal 70	0%	24,775 (24,750-24,801)	2391 (2,370- 2,413)	4,771 (4,749- 4,793)	755 (745- 764)	639 (630- 648)	1,006 (997- 1,016)
Stage III Rectal 70	3%	21,916 (21,894-21,939)	2,308 (2,289- 2,327)	4,469 (4,449- 4,488)	737 (729- 745)	600 (592- 608)	895 (887-903)
Stage II Colon 70	0%	14,754 (14,733-14,775)	2,183 (2,165- 2,200)	3,966 (3,948- 3,984)	686 (678- 694)	594 (587- 602)	983 (975-990)
Stage II Colon 70	3%	13,089 (13,071-13,108)	2,077 (2,061- 2,092)	3,673 (3,657- 3,689)	665 (658- 672)	552 (546- 559)	868 (861-874)
Stage II Rectal 70	0%	17,205 (17,183-17,226)	2,102 (2,084- 2,120)	4,568 (4,549- 4,586)	783 (775- 790)	754 (746- 762)	1,264 (1,257- 1,272)
Stage II Rectal 70	3%	14,851 (14,832-14,870)	2,005 (1,990- 2,021)	4,213 (4,197- 4,229)	758 (751- 764)	704 (697- 710)	1,120 (1,114- 1,127)
Stage III Colon 75	0%	29,020 (28,996-29,044)	3,784 (3,762- 3,805)	6,436 (6,415- 6,457)	835 (827- 844)	583 (574- 592)	809 (801-818)
Stage III Colon 75	3%	26,228 (26,207-26,250)	3,609 (3,590- 3,628)	5,994 (5,975- 6,012)	818 (811- 826)	547 (539- 555)	717 (709-725)
Stage III Rectal 75	0%	40,351 (40,325-40,378)	4,878 (4,853- 4,902)	6,728 (6,704- 6,752)	729 (722- 736)	593 (586- 600)	970 (963-977)
Stage III Rectal 75	3%	37,053 (37,029-37,076)	4,712 (4,690- 4,734)	6,398 (6,376- 6,419)	713 (707- 719)	558 (552- 564)	860 (854-867)
Stage II Colon 75	0%	12,434 (12,417-12,451)	2,208 (2,193- 2,222)	3,927 (3,912- 3,941)	643 (637- 649)	523 (517- 529)	883 (877-889)
Stage II Colon 75	3%	11,134 (11,119-11,150)	2,087 (2,075- 2,100)	3,614 (3,601- 3,626)	622 (616- 627)	485 (480- 491)	776 (771-781)
Stage II Rectal 75	0%	13,707 (13,689-13,725)	2,155 (2,140- 2,170)	3,524 (3,509- 3,538)	717 (711- 723)	685 (679- 691)	1,190 (1,184- 1,196)
Stage II Rectal 75	3%	12,396 (12,379-12,412)	2,058 (2,045- 2,071)	3,284 (3,271- 3,297)	696 (690- 701)	639 (633- 644)	1,051 (1,046- 1,056)

Table D.5 contains the results of the cost analyses. For each cohort, the third column gives the baseline mean-costs in 2017 US dollars (discounted at 0% and 3%). By baseline, I mean the average costs associated with treatment of recurrence. The next five columns give the incremental increases in mean costs comparing the surveillance-schedule identified by the column title, e.g., ‘CEA + 2 CT Scans’, with the previous regimen (the one to the left). In each cell, the first entry is

the point estimate, and 95% confidence intervals are given in the second row in parentheses. These quantify the precision of the point estimate with respect to Monte Carlo error.

Table D.6: Incremental Cost-Effectiveness Ratios (ICERs) in US \$ per Life-Year by Cohort

Cohort	Discount Rate	1 CT Scan (\$/LY)	CEA + 1 CT Scan (\$/LY)	CEA + 2 CT Scans (\$/LY)	CEA + 3 CT Scans (\$/LY)	CEA + 5 CT Scans (\$/LY)
Stage III Colon 60	0%	22,395 (22,073-22,717)	25,950 (25,664-26,236)	29,553 (28,159-30,947)	51,083 (45,680-56,486)	97,775 (85,116-110,434)
Stage III Colon 60	3%	34,736 (34,265-35,207)	39,797 (39,343-40,251)	48,521 (46,080-50,962)	93,112 (80,785-105,439)	190,945 (155,758-226,132)
Stage III Rectal 60	0%	20,089 (19,729-20,449)	22,428 (22,147-22,709)	33,083 (31,284-34,882)	79,634 (66,487-92,781)	101,446 (89,625-113,267)
Stage III Rectal 60	3%	30,962 (30,448-31,476)	34,122 (33,690-34,554)	55,703 (52,381-59,025)	173,096 (130,574-215,618)	190,504 (160,185-220,823)
Stage II Colon 60	0%	17,863 (17,611-18,115)	20,399 (20,178-20,620)	29,941 (28,501-31,381)	55,534 (49,760-61,308)	102,554 (91,380-113,728)
Stage II Colon 60	3%	26,857 (26,509-27,205)	30,317 (29,989-30,645)	48,514 (46,059-50,969)	95,290 (83,705-106,875)	180,852 (155,777-205,927)
Stage II Rectal 60	0%	15,922 (15,673-16,171)	18,227 (18,053-18,401)	26,165 (25,160-27,170)	36,071 (34,080-38,062)	54,383 (51,872-56,894)
Stage II Rectal 60	3%	24,062 (23,722-24,402)	27,078 (26,822-27,334)	41,689 (40,046-43,332)	59,043 (55,447-62,639)	89,047 (84,233-93,861)
Stage III Colon 65	0%	27,884 (27,599-28,169)	31,865 (31,601-32,129)	39,360 (37,492-41,228)	91,466 (77,363-105,569)	140,869 (120,533-161,205)
Stage III Colon 65	3%	41,061 (40,649-41,473)	46,079 (45,681-46,477)	62,115 (58,895-65,335)	183,271 (142,130-224,412)	258,245 (205,593-310,897)
Stage III Rectal 65	0%	25,610 (25,270-25,950)	28,602 (28,331-28,873)	45,578 (43,268-47,888)	119,069 (99,529-138,609)	228,835 (186,206-271,464)
Stage III Rectal 65	3%	37,555 (37,073-38,037)	41,299 (40,896-41,702)	74,079 (69,849-78,309)	257,907 (191,348-324,466)	520,557 (351,168-689,946)
Stage II Colon 65	0%	22,102 (21,875-22,329)	25,771 (25,554-25,988)	40,516 (38,739-42,293)	84,319 (75,099-93,539)	169,441 (148,173-190,709)
Stage II Colon 65	3%	31,650 (31,343-31,957)	36,302 (35,995-36,609)	61,892 (59,034-64,750)	141,704 (122,932-160,476)	299,119 (248,646-349,592)
Stage II Rectal 65	0%	20,419 (20,189-20,649)	23,284 (23,111-23,457)	38,644 (37,026-40,262)	49,815 (47,047-52,583)	84,670 (80,166-89,174)
Stage II Rectal 65	3%	29,220 (28,912-29,528)	33,044 (32,797-33,291)	60,491 (57,786-63,196)	78,699 (73,787-83,611)	135,904 (127,190-144,618)
Stage III Colon 70	0%	36,453 (36,099-36,807)	39,913 (39,574-40,252)	53,624 (51,207-56,041)	140,928 (117,349-164,507)	248,938 (195,007-302,869)

Stage III Colon 70	3%	50,230 (49,727- 50,733)	53,925 (53,450- 54,400)	79,364 (75,449- 83,279)	246,398 (190,883- 301,913)	493,829 (321,322- 666,336)
Stage III Rectal 70	0%	<u>35,757</u> (35,287- 36,227)	39,535 (39,126- 39,944)	63,090 (59,644- 66,536)	179,517 (144,359- 214,675)	344,639 (260,179- 429,099)
Stage III Rectal 70	3%	49,832 (49,154- 50,510)	53,968 (53,384- 54,552)	93,805 (88,168- 99,442)	362,823 (251,917- 473,729)	652,010 (407,408- 896,612)
Stage II Colon 70	0%	27,631 (27,354- 27,908)	30,675 (30,407- 30,943)	54,416 (51,759- 57,073)	122,290 (106,009- 138,571)	262,293 (215,672- 308,914)
Stage II Colon 70	3%	37,191 (36,827- 37,555)	40,668 (40,311- 41,025)	79,240 (75,123- 83,357)	182,423 (154,786- 210,060)	446,348 (338,006- 554,690)
Stage II Rectal 70	0%	26,307 (26,012- 26,602)	28,786 (28,558- 29,014)	49,203 (47,160- 51,246)	77,160 (71,745- 82,575)	120,954 (112,757- 129,151)
Stage II Rectal 70	3%	35,760 (35,366- 36,154)	38,608 (38,298- 38,918)	71,051 (67,961- 74,141)	113,514 (104,689- 122,339)	185,499 (170,237- 200,761)
Stage III Colon 75	0%	46,299 (45,858- 46,740)	48,475 (48,062- 48,888)	74,854 (70,856- 78,852)	261,790 (186,368- 337,212)	8,669,429 (NA)
Stage III Colon 75	3%	59,947 (59,342- 60,552)	60,689 (60,156- 61,222)	105,836 (99,563- 112,109)	466,125 (270,646- 661,604)	Dominated
Stage III Rectal 75	0%	Dominated by Extension	41,519 [^] (41,331- 41,707)	98,718 (91,793- 105,643)	523,471 (267,922- 779,020)	6,615,091 (NA)
Stage III Rectal 75	3%	Dominated by Extension	53,804 [^] (53,550- 54,058)	139,640 (128,805- 150,475)	2,091,750 (NA)	Dominated
Stage II Colon 75	0%	35,176 (34,830- 35,522)	39,985 (39,618- 40,352)	75,128 (71,030- 79,226)	195,244 (159,317- 231,171)	444,450 (331,428- 557,472)
Stage II Colon 75	3%	44,833 (44,384- 45,282)	49,637 (49,176- 50,098)	98,763 (93,256- 104,270)	313,802 (238,430- 389,174)	669,000 (450,582- 887,418)
Stage II Rectal 75	0%	31,653 (31,324- 31,982)	34,585 (34,255- 34,915)	75,350 (71,372- 79,328)	118,064 (107,439- 128,689)	230,623 (206,488- 254,758)
Stage II Rectal 75	3%	40,445 (40,021- 40,869)	43,235 (42,820- 43,650)	104,136 (98,292- 109,980)	169,247 (151,810- 186,684)	342,643 (297,873- 387,413)

Table D.6 contains the results of the cost-effectiveness analyses. Each row gives the incremental cost-effectiveness ratios (ICERs) for a given cohort and discount rate. ICERs are presented in 2017 US dollars per incremental life-year. Each ICER represents the cost per additional life-year achieved in virtue of using the surveillance-schedule identified by the column title, e.g., ‘CEA + 2 CT Scans’ instead of the previous, less-intensive regimen (the one to the left). In each cell, the first entry is the point estimate, and 95% confidence intervals are given in the second row in parentheses. These quantify the precision of the point estimate with respect to Monte Carlo error. In the event that the incremental efficacy was negative, the intervention (more intensive

surveillance) is dominated by the comparison (less-intensive surveillance) and the cell simply says 'Dominated'. If this was not the case but the 95% CI for the ICER included negative values, a CI is not given.

^ = this ICER was recalculated because it dominated by extension the next least effective intervention.

Table D.7: Sensitivity Results: Incremental Cost-Effectiveness Ratios (ICER)

Analysis	1 CT Scan (\$/LY)	CEA + 1 CT Scan (\$/LY)	CEA + 2 CT Scans (\$/LY)	CEA + 3 CT Scans (\$/LY)	CEA + 5 CT Scans (\$/LY)
Discount Rate = 0%					
Baseline	27,884 (27,599-28,169)	31,865 (31,601-32,129)	39,360 (37,492-41,228)	91,466 (77,363-105,569)	140,869 (120,533-161,205)
Salvage Treatment Costs × 1.5	37,387 (36,935-37,839)	41,528 (41,100-41,956)	53,970 (47,334-60,606)	80,303 (58,913=101,693)	217,650 (106,006-329,294)
Salvage Treatment Costs × 0.5	16,372 (16,076-16,668)	19,130 (18,902-19,358)	26,966 (23,545-30,387)	48,455 (35,530-61,380)	152,663 (74,419-230,907)
Palliative Costs × 1.5	20,359 (19,910-20,808)	24,351 (24,022-24,680)	29,144 (25,010-33,278)	53,228 (38,496-67,960)	171,663 (83,227-260,099)
Palliative Costs × 0.5	33,402 (33,020-33,784)	36,307 (35,950-36,664)	51,782 (45,529-58,035)	75,517 (55,591-95,443)	198,650 (97,029-300,271)
Salvage Continuing Costs = Stage IV	60,748 (60,137-61,359)	66,055 (65,478-66,632)	83,282 (74,278-92,286)	118,779 (89,767-147,791)	283,500 (143,351-423,649)
Salvage Continuing Costs = Stage III × 0.5	19,722 (19,396-20,048)	22,787 (22,512-23,062)	31,396 (27,213-35,579)	52,876 (38,210-67,542)	164,400 (78,986-249,814)
Cost of CT Scan × 2.0	31,140 (30,747-31,533)	31,455 (31,127-31,783)	42,983 (37,704-48,262)	67,786 (49,754-85,818)	194,000 (94,485-293,515)
Cost of CT Scan × 0.5	25,076 (24,740-25,412)	29,354 (29,045-29,663)	32,289 (28,363-36,215)	49,703 (36,713-62,693)	135,763 (66,884-204,642)
More Expensive Positive CT Scan Results	27,273 (26,918-27,628)	30,876 (30,552-31,200)	41,117 (36,063-46,171)	65,503 (48,077-82,929)	188,700 (91,910-285,490)
More Expensive Clinical Visits	28,598 (28,230-28,966)	40,219 (39,795-40,643)	40,352 (35,397-45,307)	64,297 (47,210-81,384)	186,125 (90,690-281,560)
5-FU for Unresectable Recurrence	31,986 (31,657-32,315)	35,030 (34,731-35,329)	43,802 (39,391-48,213)	62,468 (49,521-75,415)	135,281 (92,148-178,414)
Salvage Mortality Rate = 5%	27,727 (27,341-28,113)	29,960 (29,635-30,285)	42,114 (36,375-47,853)	82,257 (53,396-111,118)	156,326 (90,587-222,065)

Table D.7 gives the results of the sensitivity analyses using a discount rate of 0%. ICERs are reported in 2017 US dollars per additional life-year. The first row (Baseline) gives the reference-case ICERS for 65-year-old stage III colon cancer patients. Each subsequent row gives the

ICERS (and 95% CIs) for different univariate sensitivity analyses. The first four involved multiplying the assumed cost(s) by a factor of 1.5 or 0.5. The 5th row in each section reports the results when the continuing costs among patients treated with salvage treatment were assumed to be the same as for patients with unresectable recurrence. For the latter patients, the model used an estimate based on stage IV patients. In the reference-case analysis, the model used an estimate for continuing costs following salvage treatment based on stage III patients. The row immediately below reports the results using continuing costs equal to 50% of those of stage III patients. The next two rows present the results using an inflated and deflated cost for a CT scan, respectively. In the following two rows, analyses used larger values for the cost of a true-positive (x10) and false-positive (2x) CT scan result and an increased cost of routine clinical visits, respectively. Finally, the last two rows give the results for an analysis done assuming all (65-year old stage III colon cancer) patients with unresectable recurrence are treated with 5-FU rather than FOLFOX and for an analysis done using a salvage mortality rate of 5% rather than 2%. To determine the impact variation in a given parameter had on the ICERs, the ICERs from the appropriate row should be compared to the baseline row.